HUMBOLDT-UNIVERSITÄT ZU BERLIN
INSTITUT FÜR BIBLIOTHEKS- UND INFORMATIONSWISSENSCHAFT

BERLINER HANDREICHUNGEN
ZUR BIBLIOTHEKS- UND
INFORMATIONSWISSENSCHAFT

HEFT 227

DAS BRADFORD LAW OF SCATTERING

EMPIRISCHE ÜBERPRÜFUNG UND NUMERISCHE MODELLIERUNG

VON
FRANK JOSEF NOBER

DAS BRADFORD LAW OF SCATTERING

EMPIRISCHE ÜBERPRÜFUNG
UND
NUMERISCHE MODELLIERUNG

VON
FRANK JOSEF NOBER

Berliner Handreichungen zur
Bibliotheks- und Informationswissenschaft

Begründet von Peter Zahn
Herausgegeben von
Konrad Umlauf
Humboldt-Universität zu Berlin

Heft 227

Nober, Frank Josef

Das Bradford Law of Scattering : Empirische Überprüfung und numerische Modellierung von Frank Josef Nober. - Berlin : Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin, 2008. - 69 S. - (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft ; 227)

ISSN 14 38-76 62

Abstract:

In dieser Arbeit wird das Bradford Law of Scattering, das die Verteilung von Artikeln zu einem gegebenen Thema auf Fachzeitschriften beschreibt, untersucht. Obgleich diese Gesetzmäßigkeit im Rahmen der Bibliometrie als eine der wichtigsten gelten kann, ist sie verallgemeinert als ein Beispiel des Phänomens der Clusterbildung aufzufassen, wie es im Rahmen der statistischen Physik, der Physik der komplexen Systeme bzw. der Ökono- und Soziophysik seit langem bekannt und umfangreich mathematisch untersucht ist. Die empirische Gültigkeit des Bradford Law of Scattering konnte anhand von Stichproben aus dem Web of Science und OAIster im Prinzip bestätigt werden. Es wurden weiterhin zwei charakteristische Abweichungen zwischen den empirisch gewonnenen Stichproben und der Idealform des Gesetzes identifiziert. Diese Abweichungen konnten mit Hilfe von Simulationen erklärt werden. Abschließend wurde untersucht, welche Maße zur Beschreibung von Konzentration und Diversität, wie sie im Prinzip durch das Bradford Law of Scattering beschrieben werden, geeignet sind.

Diese Veröffentlichung geht zurück auf eine Master-Arbeit im postgradualen Fernstudiengang Master of Arts (Library and Information Science) an der Humboldt-Universität zu Berlin im SS2007.

Online-Version: http://www.ib.hu-berlin.de/~kumlau/handreichungen/h227/

Inhaltsverzeichnis

Kapitel 1

Einleitung

Bibliometrie kann als die quantitative Untersuchung der wissenschaftlichen Kommunikation mittels statistischer Methoden bezeichnet werden. Sie ist zugleich Teilgebiet der Scientometrie (also der Wissenschaft von der Vermessung der Wissenschaft – auch als Wissenschaftswissenschaft bezeichnet) als auch der Info(r)metrie, die ihrerseits dem Namen nach wiederum die „vermessende“ Subdisziplin der Informationswissenschaft ist. Erste bibliometrische Ansätze gehen etwa auf [*Hulme*, 1923] zurück. Sprach dieser noch von „statistical bibliometry“, so prägte [*Ranganathan*, 1969] den Begriff „librametry“. Beide Begriffe sind heute jedoch kaum noch zu finden. Durchsetzen konnte sich erst die von [*Otlet*, 1934] und später von [*Pritchard*, 1969] begründete Bezeichnung „bibliometrics“. Nach Pritchard beschreibt „bibliometrics . . . the application of mathematical and statistical methods to books and other media of communication“. Weiter gefasst als der Begriff der Bibliometrie ist der Begriff „Info(r)metrie“. Dieser im deutschsprachigen Raum von [*Nacke*, 1979] geprägte Begriff wird als Überschneidungsbereich zwischen den Disziplinen Mathematik und Informationswissenschaft angesehen. Hierbei geht es also weit allgemeiner um sämtliche Bereiche der Informationswissenschaft, die einer Vermessung zugänglich sind. In neuerer Zeit sind hier zuerst und vor allem die Aspekte zu nennen, die sich unter den Begriffen Cybermetrics, web metrics bzw. Internetometrie subsumieren lassen [*Mckiernan*, 2005]; [*Thelwall*, 2006]. Gemeint ist damit nichts weniger als die „Vermessung“ des Internets in Bezug auf Anzahl und Verlinkung von Webseiten - die sich daraus ergebenden Strukturen – oder auch das Verhalten von Internetnutzern. Im Gegensatz dazu beschränkt sich der Forschungsbereich der Bibliometrie nach Pritchard auf Medien, die zur Kommunikation (zwischen Wissenschaftlern) dienen.

Wissenschaftliche Kommunikation im hier gebrauchten Sinn, findet heute überwiegend mittels Veröffentlichungen in wissenschaftlichen Fachzeitschriften statt. Diese Publikationsform findet ihre Verbreitung wiederum fast ausschließlich an wissenschaftlichen Bibliotheken, wobei an erster Stelle die Universitäts - und Hochschulbibliotheken genannt werden müssen. Damit motiviert sich das immer stärker werdende Interesse der Bibliotheks- und Informationswissenschaft an bibliometrischen Fragestellungen [*Stein-Arsic*, 2003]; [*Ball*, 2004]. Zusätzlich zeigt der moderne Forschungsbetrieb selbst ein wachsendes Interesse an bibliometrischen Fragestellungen. Können diese doch dazu herangezogen werden, das zunehmende Bedürfnis nach Evaluation, Vergleich und Bewertung wissenschaftlicher Leistungsfähigkeit von Hochschulen, Instituten und ein-

zelnen Wissenschaftlern zu befriedigen. Hier bietet sich auch für Bibliotheken im wissenschaftlichen Umfeld ein neues Betätigungsfeld an, das darin besteht, bibliometrische Serviceangebote für Wissenschaft, Forschung, aber auch für politische Entscheidungsträger bereit zu halten. Damit ist jedoch nur ein Bereich für die Anwendung bibliometrischer Methoden genannt. Ein zweiter und mindestens ebenso wichtiger Punkt besteht in der Erschließung und Nutzbarmachung von Information in Form von Metainformation über wissenschaftliche Veröffentlichungen. Dieser Bereich ist seit Bestehen der ersten Bibliotheken Kernaufgabe und vornehmliches Ziel allen bibliothekarischen Bestrebens. So ist die Aufgabe von Bibliothekskatalogen seit alters her an sich die, Sammlungen und Bestände, die vermöge Umfangs und Vielfalt dem menschlichen Auffassungsvermögen entzogen sind, geordnet und übersichtlich abzubilden. Die so gewonnene Ordnung, die sich an der Metainformation (formale Titelaufnahme bzw. sachliche Erschließung) und nicht an der Primärinformation orientiert, kann dabei in vielfacher Weise manifest werden [*Löffler*, 2005]. So wie also die klassische bibliothekarische, respektive dokumentarische Tätigkeit den Inhalt eines Buches, Dokuments oder einer medialen Einheit in Form der Titelaufnahme oder eines bibliographischen Eintrags kondensiert darstellt, so vermag die Bibliometrie – ansetzend an diesem Pool von Metainformationen – ihrerseits eine weitere Reduktion des ursprünglichen Informationsgehaltes zugunsten neu gewonnener Informationen über Strukturen dieses Informationspools zu leisten. Bedingt durch das enorme Wachstum wissenschaftlicher Informationen setzt sich die Einsicht über den Nutzen derartiger, strukturierter Metainformation mehr und mehr durch. Gerade in Zeiten, in denen elektronische Information in einer bisher nicht gekannten Fülle weltweit verfügbar ist, werden Sinn und Notwendigkeit einer solchen Strukturierung von Informationsressourcen manifest.
Der Begriff „Wissensgesellschaft“ erfreut sich derzeit großer Beliebtheit. Ob unsere Gesellschaft aber nun eine Wissens - oder nicht vielmehr nach [*Umstätter*, 1997] eine Wissenschaftsgesellschaft ist, sei hier anheimgestellt. In jedem Fall aber befinden wir uns in einem Zeitalter, in dem Information, Wissen und Wissenschaft eine nie gekannte Bedeutung für alle Bereiche des öffentlichen, privaten und vor allem des wissenschaftlichen Lebens gewonnen haben [*BMBF*, 2002a]; [*BMBF*, 2002b]. Diesem Faktum muss sich auch die Bibliotheks- und Informationswissenschaft stellen. In einem Umfeld, in dem Informationsangebote und Quellen zunehmend vielfältiger und umfangreicher werden, steht die Bibliotheks- und Informationswissenschaft in der Pflicht, Methoden und Theorien zu entwickeln, um der Herausforderung der oft zitierten Informationsflut Herr zu werden. Nur scheinbar ergibt sich ein Paradoxon, wenn man bedenkt, dass sich für Bibliotheken (vor allem, aber nicht nur im universitären Umfeld) aus der Problematik der Informationsflut ein zweites, anscheinend entgegengesetztes Problem ergibt: das des Informationsmangels. Informationsmangel in der Form, dass von Wissenschaftlern gewünschte, aus den Etats der Universitätsbibliotheken zunehmend jedoch nicht mehr bezahlbare Zeitschriften, nicht abboniert werden können. Hier könnte eine praxisnah angewandte Bibliometrie helfen, das Informationsbedürfnis einer Hochschule bzw. ihrer einzelnen Lehrstühle klarer zu definieren und die stetig knapper werdenden Etats effektiver einzusetzen.
Als die drei Schwerpunkte, die sich derzeit im Fokus bibliometrischer Fragestellungen finden, sind zu nennen:
1) Messung der Resonanz auf Veröffentlichungen von Forschergruppen bzw. einzelnen

Wissenschaftlern und Bestimmung ihres Gewichts in der Wissenschaftsgemeinschaft.
2) Analyse von thematischen Trendentwicklungen: Das Auftreten von thematischen "hot spots" in der Vergangenheit kann analysiert werden und zur Vorhersage neuer Themenschwerpunkte anhand der Veröffentlichungszahlen genutzt werden.
3) Bestimmung der Interdisziplinarität von wissenschaftlichen Teildisziplinen aufgrund der Verteilung von Veröffentlichungen auf die Menge aller wissenschaftlichen Zeitschriften. Allgemeiner aufgefasst ist davon auszugehen, dass mit Hilfe mathematischer Methoden der wissenschaftliche Publikationsprozess analysiert und hinsichtlich auftretender Strukturen untersucht werden kann.
Der letztgenannte Punkt ist Thema der vorliegenden Arbeit. Ein aktueller und umfangreicher Überblick über die Entwicklungen im Bereich Bibliometrie und Informetrie sowie eine ausführliche Darstellung der mathematischen Techniken findet sich bei [*Egghe*, 2005].

Geht es beim Thema der vorliegenden Arbeit darum, Methoden des quantitativen, wissenschaftlichen Arbeitens auf die Wissenschaft selbst anzuwenden, so soll nicht verschwiegen werden, dass dieses Vorgehen gleich in mehrfacher Weise kritisch betrachtet werden kann. Zum einen ist die Methode selbst betroffen. Wird z.B. der Prozess des Veröffentlichens wissenschaftlicher Artikel als reiner Zufallsprozess, der von gewissen systemimmanenten Wahrscheinlichkeiten gesteuert wird (dieses Vorgehen wird wesentlicher Inhalt dieser Arbeit sein), betrachtet, so bleiben wesentliche Faktoren, die das soziale Gefüge „Wissenschaft" beschreiben, unberücksichtigt. Diese Kritik muss seitens der Bibliometrie angenommen werden, ist jedoch nicht wirklich fundamental. Bei geeigneter Konzeption der mathematischen Modelle lassen sich prinzipiell alle Einflüsse – wenn auch nur auf statistische Weise – berücksichtigen. Die entscheidende Frage ist hier eher, wie stark man ein System abstrahieren kann und will. Die prinzipielle Möglichkeit, komplexe soziologische Prozesse mit Hilfe der Methoden der statistischen Physik zu beschreiben, ist im Bereich der Soziophysik hinlänglich belegt [*Mainzer*, 1999]; [*Schweitzer*, 2003].
Ein zweiter Kritikpunkt richtet sich weniger gegen bibliometrische Analysen an sich als vielmehr gegen ihre von manchen [*Fröhlich*, 1999]; [*Naumann*, 2006] als verheerend angesehene Rückwirkung auf die Wissenschaft selbst. In dem Moment, so wird bemängelt, wo bestimmte Methoden erdacht und Indikatoren definiert werden, um etwa die wissenschaftliche Leistungsfähigkeit von einzelnen Wissenschaftlern oder gar ganzen wissenschaftlichen Institutionen zu beurteilen und gegeneinander ins Verhältnis zu setzen, werden sich Wissenschaftler und letztlich die gesamte wissenschaftliche Gemeinschaft in einer Weise anpassen, die ihnen ein besonders gutes ranking verspricht. Das System (Wissenschaft) reagiert, wenn auch mit einer gewissen Trägheit und zeitlichen Verzögerung, auf die durchgeführte Messung, die – und dies schließt den kausalen Zusammenhang –, so sie etwa zur Bestimmung von Leistungsindikatoren durchgeführt wird, ja zur Steuerung wissenschaftspolitischer Entscheidungen dienen soll. Eine genaue Diskussion der Aus- bzw. Rückwirkungen bibliometrischer Methoden und Aussagen auf den wissenschaftlichen Betrieb ist im Rahmen dieser Arbeit sicherlich nicht möglich; sie würde leicht eine eigene eigenständige Forschungsarbeit füllen. Das in den Wissenschaften zum geflügelten Wort gewordene „publish or perish" beschreibt die Misere (die eine falsch verstandene Bibliometrie mit sich gebracht hat)

jedoch auf drastische Art und Weise. Publikationszahl und Informationsgehalt bzw. Qualität der Forschung sind nicht zwangsläufig proportional – genau dies spiegelt jedoch eine falsch verstandene und falsch angewandte Bibliometrie vor. Hier besteht aber gerade auch ein Betätigungsfeld bibliometrischer Forschung: Nicht mehr einfach nur zu zählen, sondern auch inhaltliche Aspekte einfließen zu lassen, um auf diesem Wege bewerten zu können, ob eine Veröffentlichung auch tatsächlich neue Information enthält. Dieser Weg, der derzeit sicherlich noch Zukunftsmusik ist, beschreibt eine mögliche Zukunft der Bibliometrie.
Bedient sich die Bibliometrie statistischer Methoden, so ist damit zunächst und vor allem der Umgang mit großen Datenmengen verbunden. Die schnelle und einfache Verfügbarkeit großer Datenmengen wurde erst mit dem Aufkommen elektronischer, digitaler Datenbanken möglich. Damit lässt sich auch erklären, warum die Bibliometrie, obgleich sie ihre Wurzeln – wie oben erwähnt – in der ersten Hälfte des letzten Jahrhunderts hat, erst in jüngerer Zeit in der Praxis tatsächlich durchführbar geworden ist. Der Begriff „groß“ muss in diesem Zusammenhang zunächst als relativ unbestimmt im Raume stehen bleiben. Ob eine Stichprobe an Daten – gleich welcher Natur – als „groß genug“ zur Anwendung statistischer Methoden gelten kann, ist ohne genaue Kenntnis der Grundgesamtheit, d.h. der Menge aller Daten, die das betrachtete System vollständig beschreiben würde, a priori nicht zu bestimmen. Diese genaue Kenntnis ist in der Regel bei empirischen Daten nicht gegeben, sondern muss durch heuristische Annahmen erlangt werden. Ohne eine tiefer gehende Diskussion dieser Problematik zu führen, lauten die relevanten Fragen zur Anwendung statistischer Methoden im Rahmen der vorliegenden Arbeit:
1) Bilden das Web of Science bzw. OAIster als bibliographische Datenbank bzw. als Metasuchmaschine für frei zugängliche online Repositorien den Publikationsmarkt in repräsentativer Weise ab oder nicht?
2) Lassen die Teilmengen an Publikationen, die sich in diesen beiden Datenbanken zu einem thematischen Suchbegriff finden lassen, belastbare Aussagen über die Verteilung von Publikationen zu einem bestimmten Thema auf Zeitschriften bzw. online Repositorien zu?
Diese beiden Fragen müssen, damit das Programm dieser Arbeit überhaupt sinnvoll ist, a priori mit „Ja“ beantwortet werden. Es gibt keinen harten Beweis, der hierfür ins Felde geführt werden könnte, sondern eher die Erfahrung des Praktikers und die Betrachtung gewisser Kenngrößen.

Das Bradford Law of Scattering

Das Gesetz von Bradford (geläufiger unter dem Namen: Bradford Law of Scattering) kann sicherlich als das wichtigste unter den bisher gefundenen bibliometrischen Gesetzmäßigkeiten angesehen werden und wird im Weiteren Gegenstand dieser Arbeit sein. Es wurde 1934 von Samuel C. Bradford [*Bradford*, 1934] formuliert und beschreibt in statistischer Weise die Verteilung von Literatur zu einem bestimmten Thema über verschiedene Fachzeitschriften. Nach dem Bradford Law of Scattering (von nun an kurz mit BLS bezeichnet) findet sich immer die gleiche Anzahl von Zeitschriftenartikeln in Gruppen von Fachzeitschriften, deren Mächtigkeit sich verhält wie: $n^0 : n^1 : n^2 : \ldots$ wobei n ein vom Fachgebiet abhängiger Parameter ist [*Umstätter*, 2005]. Vereinfacht besagt dieses Gesetz, dass es für jedes Fach eine Kernzeitschrift gibt, in der ein wesentlicher

Teil aller Artikel zu einem Thema zu finden ist. Der Rest der Artikel verteilt sich dann zunehmend breiter gestreut auf andere Zeitschriften. In dieser Formulierung wird die Verwandtschaft zur Pareto-Regel ersichtlich, nach der man mit 20% des Aufwands 80% des Erfolgs erzielt. Analog erhält man also durch Sichtung von 20% der wichtigsten Zeitschriften 80% der relevanten Artikel zu einem Thema. Diese ursprüngliche Formulierung des BLS zeigt stark qualitativen Charakter. Es wird wesentlicher Teil dieser Arbeit sein, die Formulierung des BLS sowie dessen Verwandtschaft zu anderen aus allen Bereichen der Wissenschaften bekannten charakteristischen Verteilungen aufzuzeigen und auf einen fundierten, mathematischen Formalismus zu stützen.

1.1 Überblick über die Arbeit

Nach dieser Einleitung soll nun ein Überblick über das Programm dieser Arbeit gegeben werden: In Kapitel 2 wird das BLS, streng angelehnt an die Beschreibung anderer Prozesse aus dem Bereich der Physik der komplexen Systeme, mathematisch formuliert. Es wird der für den ersten Teil der Arbeit notwendige mathematische Apparat eingeführt. Kapitel 3 folgt mit der Vorstellung des empirischen Fundaments dieser Arbeit. Das Web of Science sowie die Metasuchmaschine OAIster werden jeweils beschrieben und die prinzipiellen Unterschiede, die für die Interpretation der Stichproben von Belang sind, dargelegt. Anschließend werden die Stichproben dargestellt und diskutiert. Kapitel 4 widmet sich dann speziell den mathematischen Grundlagen der Simulation stochastischer Prozesse. Das numerische Modell wird vorgestellt und sein Konzept begründet. Es werden qualitative Ergebnisse sowie umfangreiche Sensitivitätstests vorgestellt, bevor der Vergleich zwischen empirischen Daten und numerischen Ergebnissen vollzogen wird. In Kapitel 5 werden alternative Maße zur Bestimmung von Diversität und Konzentration in Verteilungsfunktionen vorgestellt, angewandt und kurz diskutiert. Die Arbeit endet schließlich mit einer allgemeinen Diskussion sowie einem Ausblick.

Kapitel 2

Mathematische Grundlagen

Potenzgesetze in den unterschiedlichsten Varianten und Ausprägungen findet man in ganz unterschiedlichen Bereichen der quantitativen Wissenschaften. Genannt seien hier nur einige ausgewählte Beispiele: Die Größenverteilung von Wolken (genauer: flacher konvektiver Wolken) folgt diesem Gesetz ebenso, wie etwa die Verteilung des Einkommens auf die Bevölkerung eines Landes oder die Intensität von Sonneneruptionen. Bestimmt man die Verteilungsfunktion der Bevölkerungszahlen von Städten eines Landes (oder auch den Ländern auf der Erde), so erhält man ebenso eine Potenzfunktion wie für die Größenverteilung von Dateien in Betriebssystemen von Computern. Erdbeben und Lawinen, Wasserwellen und Waldbrände, die Verlinkung von Websites und die Häufigkeit von Namen, die Größe der Krater auf dem Mond, wie auch die Intensität der Kriege zwischen 1816 und 1980, alle diese Prozesse folgen einer Potenzgesetzverteilung. Als unvollständige und beispielhafte Literatursammlung wird auf folgende Quellen verwiesen: [*Auerbach*, 1913]; [*Zipf*, 1949]; [*Gutenberg*, 1944]; [*Neukum*, 1994]; [*Lu*, 1991]; [*Crovella*, 1996]; [*Roberts*, 1998]; [*Estoup*, 1916]; [*Zanette*, 2001]; [*Lotka*, 1926]; [*de Solla Price*, 1965]; [*Adamic*, 2000]; [*Pareto*, 1896]. Als zwei herausragende und erstklassige Überblicksartikel zu diesem Thema seien die beiden Arbeiten [*Newman*, 2005] und [*Mitzenmacher*, 2003] genannt. In diesen beiden Quellen wird eine sehr viel genauere und umfangreichere Darstellung des Themenkreises gegeben als dies hier geschehen kann. Die in der vorliegenden Arbeit gegebene Darstellung orientiert sich weitestgehend an der Nomenklatur von [*Newman*, 2005].

Die Vielfalt der Prozesse und Erscheinungen, die ein Potenzgesetzverhalten zeigen, macht deutlich, dass es sich hierbei um ein fundamentales Verhalten handeln muss. Gleichzeitig ist es ein Verhalten, das an Systeme mit einer großen Zahl von Elementarobjekten gekoppelt ist, die in einer in der Regel nicht immer bekannten und möglicherweise komplexen Art und Weise miteinander in Wechselwirkung stehen. Zur Beschreibung solcher Systeme hat sich im Rahmen der statistischen Physik der notwendigerweise interdisziplinäre Teilbereich der „Komplexen Systeme“ herausgebildet [*Mainzer*, 1999]. Als prominenteste Vertreter der nicht physikalischen Anwendungen haben sich hier die Ökonophysik und die Soziophysik etabliert [*Schweitzer*, 2003].

Betrachtet man nun das BLS, so sind die entsprechenden Elementarobjekte Wissenschaftler, die ihre Arbeiten in einer bestimmten Fachzeitschrift veröffentlichen. Die

Entscheidungsprozesse, die zu der Wahl einer bestimmten Zeitschrift führen, mögen im jeweiligen Einzelfall aus subjektiver Sicht des handelnden Wissenschaftlers in hohem Maße durch eine Vielzahl spezieller und individueller Einflüsse bestimmt sein. Die Auswahl einer Fachzeitschrift für eine Veröffentlichung gehört sicherlich zu den wichtigen strategischen Entscheidungen, die ein Wissenschaftler zu treffen hat – in der Summe der gesamten wissenschaftlichen Gemeinschaft ergibt sich jedoch ein charakteristisches Muster. Dies ist ähnlich wie im Straßenverkehr: Jeder Autofahrer wird stets überzeugt sein, hoch individuelle Entscheidungen bei der Wahl von Fahrspur, Tempo etc. zu treffen. Dennoch lässt sich das statistische Verhalten und die Dynamik des Autoverkehrs mit Methoden der statistischen Physik recht gut berechnen [*Helbing*, 1999].

Das BLS wird in der Literatur oft in der Form angegeben, dass sich immer die gleiche Menge an Literatur (in der Regel sind damit Zeitschriftenartikel gemeint) auf verschiedene, thematisch geordnete Zeitschriften wie $n^0 : n^1 : n^2 : n^3 ...$ verteilt. Als Zahlenbeispiel mit $n = 10$ hieße dies also, dass sich eine bestimmte Anzahl von Artikeln in der einen wichtigen Kernzeitschrift finden. Um die gleiche Anzahl von Artikeln in den nächstwichtigen Zeitschriften zu finden, benötigt man 10 Zeitschriften. Für den nächsten Schritt 100 Zeitschriften und so weiter. Es ist augenfällig, dass diese Art der Formulierung nicht durch mathematische Eleganz besticht. Es handelt sich vielmehr um das mühsam in Worte gefasste Phänomen der Clusterbildung. Wie wichtig dieses Phänomen in vielen Bereichen von Naturwissenschaft und Technik, aber auch bei gesellschaftlichen Prozessen ist, kann hier nicht im Detail dargelegt werden. Stattdessen sei noch einmal auf die obige Quellensammlung verwiesen. In jedem Fall ist Clusterbildung also kein Spezifikum der Bibliometrie. Im Gegenteil: Man wird beobachten können, dass bei Prozessen, bei denen eine mehr oder minder komplexe Wechsel- und Rückwirkung zwischen Objekten zu einer Clusterbildung führt, jeweils im Detail natürlich grundverschiedene, aber dennoch analoge Grundtendenzen beteiligt sind. Diese Grundtendenzen spiegeln immer zwei widerstreitende Kräfte wieder: Eine Tendenz zur Konzentration (also zur Bildung von Clustern) und eine Tendenz zur Diversifizierung (ohne die sich im Grunde immer nur die triviale Verteilung von nur einem einzigen Cluster einstellen würde). Die resultierende Verteilung ist dann jeweils eine Art Gleichgewichtsverteilung für die beteiligten Kräfte. Im vorliegenden Fall wäre die Tendenz zur Konzentration sicherlich dadurch motiviert, dass sich in der Fachwelt anerkannte und wichtige Zeitschriften herauskristallisieren, bei denen die veröffentlichten Artikel ein besonders hohes Gewicht haben. Jeder Wissenschaftler eines mehr oder minder eng umgrenzten Gebiets wird bestrebt sein, in diesen Zeitschriften zu veröffentlichen. Andererseits wird sich aufgrund der stets vorhandenen interdisziplinären Tendenzen eines jeden Fachgebiets immer auch die Notwendigkeit zur Diversifizierung ergeben. So wird sich jedes Gebiet der Wissenschaften immer in Teilgebiete mit unterschiedlichen Schwerpunkten unterteilen. Diese zwar keinesfalls bewiesene, dennoch aber plausible Annahme wird in Kapitel 4 wesentlich in die Konzeption des Modells zur Simulation des BLS einfließen.

Im Folgenden soll jedoch zunächst der Sachverhalt am Beispiel der Verteilung der 1000 größten deutschen Städte noch einmal veranschaulicht werden. Anhand dieses zugänglichen Beispiels wird der in der Mathematik und Physik exakt definierte ma-

thematische Formalismus zur Beschreibung solcher Verteilungen vorgestellt.

2.1 Das Potenzgesetz, Verteilungsfunktion und Wahrscheinlichkeitsdichte

Um den Formalismus einer Potenzgesetzverteilung möglichst anschaulich zu motivieren, soll er hier anhand des Beispiels der 1000 größten Städte in Deutschland eingeführt werden [Quelle: http://bevoelkerungsstatistik.de/]. In Abbildung 2.1 sieht man die absolute Anzahl der 1000 größten deutschen Städte aufgetragen über ihre Einwohnerzahl. Man erkennt das starke Gefälle: Es gibt eine große Zahl an Städten mit einer geringen Einwohnerzahl und eine immer geringer werdende Anzahl an Städten bei wachsender Einwohnerzahl. Eine typische Strategie, um ein solches Diagramm, bei dem sich die Kurve allzu stark an die Koordinatenachsen anschmiegt, besser darzustellen, ist der Wechsel der Achsenskalierung von einer linearen hin zu einer logarithmischen Skalierung. Vollzieht man diesen Wechsel (Abb. 2.2), so erkennt man schon recht gut das grundlegende Verhalten der Verteilung. Es zeigt sich bei logarithmischer Skalierung beider Achsen eine relativ gut ausgeprägte Gerade.

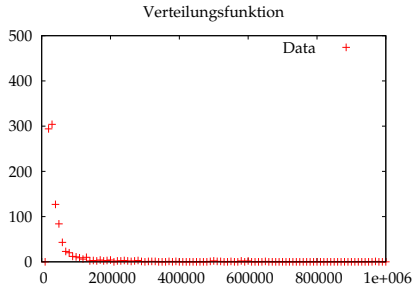


Abb. 2.1: Verteilungsfunktion: Zahl der Städte aufgetragen über deren Einwohnerzahl - linear skaliert

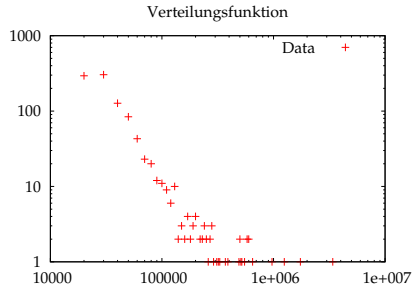


Abb. 2.2: Verteilungsfunktion: Zahl der Städte aufgetragen über deren Einwohnerzahl - logarithmisch skaliert

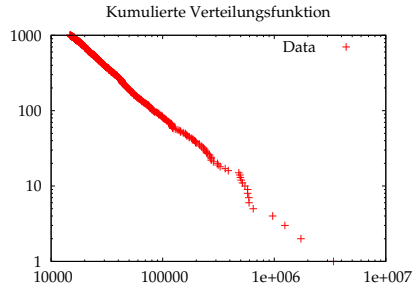


Abb. 2.3: Kumulierte Verteilungsfunktion

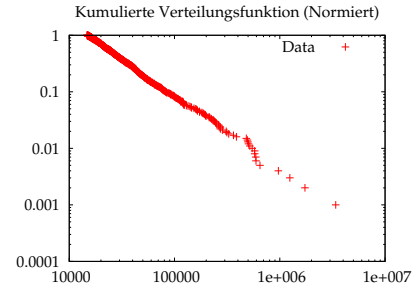


Abb. 2.4: Kumulierte Verteilungsfunktion - Normiert

Sei nun $f(x)$ die Anzahl an Städten mit einer Einwohnerzahl zwischen x und $x + dx$. Da die Verteilung logarithmisch aufgetragen eine Gerade ergibt, folgt für $f(x)$:

$$\log(f(x)) = \log(C) + (-\alpha)\log(x) \tag{2.1}$$
$$\tag{2.2}$$

Nach Anwendung der Rechengesetze für den Logarithmus ergibt sich:

$$\begin{aligned} \log(f(x)) &= \log(Cx^{-\alpha}) \\ f(x) &= Cx^{-\alpha} \end{aligned} \tag{2.3}$$
$$\tag{2.4}$$

Somit spricht man bei der Funktion $f(x)$ von einem Potenzgesetz. α ist der Exponent der Verteilung und C eine Normierungskonstante. Am vorliegenden Beispiel (Abb. 2.2) mag die Argumentation, die Verteilung folge einer Geraden, noch nicht hinreichend überzeugend sein. Dies liegt noch zum Großteil an der ungünstigen Darstellung der Daten. In der Abb. 2.2 wurden die Daten in ein festes Raster von Einwohnerzahlen aufgetragen, d.h. ein Punkt entspricht jeweils der Anzahl der Städte mit Einwohnerzahl zwischen x und $x + 10.000$ Einwohnern. Nun liegt es in der Natur des Datensatzes, dass es sehr viel mehr Städte mit Einwohnerzahlen zwischen 10.000 und etwa 1.000.000 gibt als darüber. Bei großen Werten wird die Darstellung der Werte aufgrund ihrer geringen Anzahl dadurch derart verrauscht, dass das charakteristische Verhalten fast nicht mehr zu erkennen ist. Abhilfe könnte hier ein logarithmisch anwachsendes Raster liefern, was jedoch eine recht aufwendige und wenig elegante Lösung wäre. In dieser Arbeit wird daher die alternative Darstellungsform mit Hilfe der kumulierten Verteilungsfunktion gewählt. Hierbei wird nicht die Verteilungsfunktion an sich dargestellt, sondern die Wahrscheinlichkeit $P[X \geq x]$, dass die Funktion einen Wert größer oder gleich x hat. Dieser Zusammenhang lässt sich durch das Integral über die Verteilungsfunktion beschreiben. In der allgemeinen Form lautet dies:

$$P[X \geq x] = \int_x^\infty f(z)dz = C \int_x^\infty z^{-\alpha} dz = \frac{C}{\alpha - 1} x^{-(\alpha-1)} \tag{2.5}$$

Beim Übergang von kontinuierlichen zu diskreten Variablen wird das Integral zur Summe, und der Gesamtausdruck ergibt sich zu:

$$P[X \geq x_j] = C \int_x^\infty z^{-\alpha} dz = C \sum_{i=j}^{n} x_i^{-\alpha} \tag{2.6}$$

Eine andere Darstellung zur Beschreibung des Potenzgesetzes bedient sich der Beta-Funktion $B(k, \alpha)$ und der Gamma-Funktion $\Gamma(\alpha]$. Hierbei gilt:

$$p_k = C \frac{\Gamma(k)\Gamma(\alpha)}{\Gamma(k + \alpha)} = C\, B(k, \alpha), \tag{2.7}$$

Und für große $k : B(k, \alpha) \sim k^{-\alpha}$
Damit folgt wiederum:

$$1 = C \sum_{k=1}^{\infty} B(k, \alpha) = \frac{C}{\alpha - 1}, \tag{2.8}$$

Mit $C = \alpha - 1$ folgt schließlich:

$$p_k = (\alpha - 1)B(k, \alpha). \tag{2.9}$$

Der Formulierung mit Hilfe der Beta-Funktion werden wir in Kapitel 4 bei der Herleitung des theoretischen Modells wieder begegnen.

Zurückkommend zur kumulierten Verteilungsfunktion gilt also nun: Die kumulierte Verteilungsfunktion weist eine (vom Betrag her) um 1 verringerte Steigung $-\alpha + 1$ auf. Abb. 2.3 zeigt diese kumulierte Verteilungsfunktion für unser Städtebeispiel. Abb. 2.4 zeigt die gleiche Funktion normiert und vollzieht damit den Übergang zur Wahrscheinlichkeitsverteilung. Somit kann man an Abb. 2.4 ablesen, dass etwa 10 % aller Städte eine Bevölkerungszahl von über 100.000 Einwohnern besitzen. Diese Normierung ist von Vorteil, wenn man Stichproben unterschiedlichen Umfanges miteinander vergleichen will.

Soll die Verteilung $f(x)$ eine Wahrscheinlichkeitsverteilung repräsentieren, so muss zusätzlich noch eine Normierung durchgeführt werden. Diese Bedingung liefert den Wert für die Konstante C.

$$1 = \int_{x_{min}}^{\infty} f(x)dx = C \int_{x_{min}}^{\infty} x^{-\alpha}dx = \frac{C}{1 - \alpha} \left[x^{-\alpha+1}\right]_{x_{min}}^{\infty} \tag{2.10}$$

Für $\alpha > 1$ kann die Normierung durchgeführt werden. Es folgt für C:

$$C = (\alpha - 1)x_{min}^{\alpha-1} \tag{2.11}$$

Im Folgenden soll nun noch kurz der mathematische Zusammenhang zwischen einer Potenzgesetzverteilung und der sogenannten 80/20 oder Pareto-Regel dargestellt werden.
Allgemeiner geht es hierbei um die Frage, wie stark die Verteilung konzentriert ist bzw. wie steil die Potenzfunktion abfällt. Hierzu soll zunächst der Median der Verteilung berechnet werden. Dies ist der Punkt $x_{1/2}$, der die Funktion $f(x)$ derart unterteilt, dass eine Hälfte der Werte unterhalb und eine Hälfte der Werte oberhalb von $x_{1/2}$ liegen. Auf das Beispiel der Städte übertragen bedeutet $x_{1/2}$, dass die Hälfte der betrachteten Städte eine geringere Einwohnerzahl als $x_{1/2}$ hat und die andere Hälfte mehr Einwohner hat.

$$\int_{x_{1/2}}^{\infty} f(x)dx = \frac{1}{2} \int_{x_{min}}^{\infty} f(x)dx \tag{2.12}$$

oder

$$x_{1/2} = 2^{1/(\alpha-1)} x_{min} \tag{2.13}$$

Eine andere Betrachtungsweise ist, umgekehrt zu fragen, wie viele Einwohner von den jeweiligen Hälften repräsentiert werden.

$$\frac{\int_{x_{1/2}}^{\infty} x f(x) dx}{\int_{x_{min}}^{\infty} x f(x) dx} = \left(\frac{x_{1/2}}{x_{min}}\right)^{-\alpha+2} = 2^{-(\alpha-2)/(\alpha-1)} \tag{2.14}$$

Dieser Wert, der Anteil der Bevölkerung in der Hälfte mit den größeren Städten, ist gegeben durch Gleichung 2.14.

Allgemeiner kann man nun für beliebige Bruchteile definieren:

$$W(z) = \frac{\int_{z}^{\infty} x f(x) dx}{\int_{x_{min}}^{\infty} x f(x) dx} \left(\frac{z}{x_{min}}\right)^{-\alpha+2} \tag{2.15}$$

Der Anteil der Bevölkerung, die in Städten leben, deren Einwohnerzahl den Wert x übertrifft, ist gegeben durch:

$$W = P^{(\alpha-2)/(\alpha-1)} \tag{2.16}$$

P gibt hierbei den betrachteten Bruchteil an.
Abb. 2.5 verdeutlicht nun noch genauer den Zusammenhang mit der 80/20 Regel. Aufgetragen ist hier entsprechend Gl. 2.16 $W(P)$ über P. Tatsächlich wird bei einer Potenzgesetzverteilung mit $\alpha = 2.15$ bereits ein Anteil von 80% aller Artikel zu einem gegebenen Thema in nur 20% der Zeitschriften, die überhaupt zu diesem Thema veröffentlichen, zu finden sein. Diese müssen natürlich die 20% der größten Zeitschriften sein, was durch die Definition in Gleichung 2.15 sichergestellt ist. Die Darstellungsform in Abb. 2.5 ist in der Literatur auch unter dem Namen Pareto-Kurve bekannt. Sie ist eng verwandt mit der Lorenz-Kurve, auf die noch in Kapitel 5 gesondert eingegangen wird.

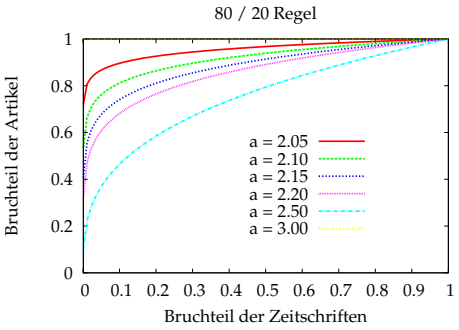


Abb. 2.5: Pareto oder 80 / 20 Regel;
Aufgetragen: W über P

Die Analogie zwischen dem Städte-Beispiel und dem BLS ergibt sich zwangsläufig: Die größte deutsche Stadt (Berlin) entspricht der Kernzeitschrift, die zu einem gegeben Thema den größten Anteil aller Artikel in sich vereinigt. So wie es nun viele Städte mit 100.000 Einwohnern und noch mehr kleinere Städte mit 10.000 Einwohnern gibt, verteilen sich auch wissenschaftliche Artikel in immer kleineren Clustern auf die Menge der Fachzeitschriften.

Kapitel 3

Empirische Daten

Um das BLS empirisch zu überprüfen, wurden für diese Arbeit insgesamt 40 Stichproben genommen. Diese wurden zu je 20 willkürlich und nicht repräsentativ gewählten Stich- bzw. Suchworten aus den beiden Quellen „Web of Science“ und „OAIster“ gezogen. Die relativ geringe Zahl an Stichproben macht deutlich, dass es sich bei dieser Untersuchung keinesfalls um eine vollständige Bestandsaufnahme und Charakterisierung der beiden Instanzen handeln kann. Diese wäre im Rahmen einer Masterarbeit weder zu leisten, noch wäre sie derzeit ohne wichtige Erkenntnisse, wie sie etwa auch innerhalb dieser Arbeit gewonnen werden, sinnvoll. Es sei jedoch schon an dieser Stelle ausdrücklich betont, dass eine solche umfassende und erschöpfende „Vermessung“ des wissenschaftlichen Publikations- und Kommunikationswesens, richtig durchgeführt, äußerst interessant wäre. Für eine tiefere Diskussion muss jedoch auf den Ausblick am Ende dieser Arbeit verwiesen werden.

Zunächst wird im Folgenden anhand der begrenzten Stichprobenzahl das BLS qualitativ überprüft. Um hierzu einerseits eine größere Menge an Daten, die eine gewisse Vielfalt aufweisen, zu erlangen, die aber andererseits nicht zu groß ist, um noch detailliert untersucht werden zu können, wurde die Anzahl von 2 mal 20 Stichproben als ein praktikabler Mittelweg zwischen „zu vielen“ und „zu wenigen“ Daten angesehen.

3.1 Quellen

Als Quellen für die Stichproben wurden die beiden Datenbanken „Web of Science“ und „OAIster“ ausgewählt. Die Bezeichnung „Datenbank“ ist sicherlich (vor allem im Fall von OAIster) verkürzt und vereinfacht, soll aber aus Gründen der sprachlichen Vereinfachung gelegentlich für beide Instanzen benutzt werden. Auf die Besonderheiten und Eigenheiten dieser Quellen wird in den nächsten beiden Abschnitten kurz eingegangen. Als Hauptgründe für die Auswahl sind sicherlich die Datenlage, insbesondere der Umfang der jeweiligen Datenbasis sowie deren Zugänglichkeit zu nennen.

3.1.1 Web of Science

Das Web of Science kann, obgleich sich zunehmend leistungsstarke Konkurrenten auf dem Markt der bibliographischen Datenbanken etablieren, wohl immer noch als das

bedeutendste Verzeichnis bibliographischer Daten (vor allem im STM-Bereich) angesehen werden. Diese Datenbank war lange Zeit die einzige multidisziplinäre Sammlung bibliographischer Daten in Verknüpfung mit deren Zitation. Dieser Science Citation Index (SCI) kann gewissermaßen als erster Versuch bezeichnet werden, ein leistungsfähiges Werkzeug für bibliometrische Analysen bereitzuhalten. Entwickelt wurde der SCI an dem 1960 von Eugene Garfield gegründeten „Institute for Scientific Information" (ISI) und gehört nun zur Thomson-Scientific Inc. Laut Thomson werden derzeit etwa 9.000 Zeitschriften im Web of Science ausgewertet. Die Diskussion, ob diese 9.000 Zeitschriften einen repräsentativen Ausschnitt des gesamten wissenschaftlichen Zeitschriftenmarktes geben, kann und soll hier nicht geführt werden. Jeder, zumindest der im Bereich der Naturwissenschaften Tätige, wird aus eigener Erfahrung die hinreichende Vollständigkeit des Web of Science für die Literaturrecherche bestätigen können. Es ist damit gleichsam im Bradfordschen Sinne eine Kerndatenbank für weite Bereiche der modernen Wissenschaften. Es sei lediglich angemerkt, dass nun nach dem Aufkommen der schon erwähnten Alternativen (namentlich sind dies vor allem googlescholar, scirus, citeseer und scopus) zum ersten Mal die Möglichkeit besteht, das Web of Science – und damit natürlich auch den SCI – mit anderen Instanzen zu vergleichen [*Pipp*, 2006]; [*Wildner*, 2006]; [*Schneider*, 2006]; [*Gorraiz*, 2006].

Das Web of Science bietet über seine Schnittstelle zum Literaturverwaltungsprogramm Endnote eine bequeme Möglichkeit, auch größere Datenmengen, d.h. Treffer auf eine Suchanfrage, auf den lokalen Rechner zu laden. Zu erwähnen ist jedoch, dass der Datentransfer ab einer Treffermenge von etwa 20.000 schließlich doch so langsam und instabil wird, dass im Weiteren bei den Suchanfragen gegebenenfalls der Suchzeitraum eingeschränkt wurde. Hierdurch wurde erreicht, dass die untersuchten Stichproben (bis auf Ausnahmen) von der Größenordnung 10.000 sind.

3.1.2 Die Metasuchmaschine OAIster

OAIster ist die derzeit größte Metasuchmaschine für Dokumente, die weltweit verteilt auf Dokumentenservern (sog. Online-Repositorien) bereitliegen. Die abgefragten Repositorien müssen hinsichtlich ihrer Erschließung durch Metadaten dem Standard der Open Archive Initiative (OAI) entsprechen. Im Sinne der OAI ist OAIster damit ein OAI Service Provider. Die Metasuchmaschine enthält mit fast 9 Millionen Datensätzen, die von ca. 700 Institutionen bereitgestellt werden, einen beeindruckenden Datenpool, der die zunehmende Bedeutung des Open-Access Gedankens widerspiegelt. Die Berücksichtigung von OAIster in der Analyse soll Aufschluss darüber geben, ob das BLS auf die Instanzen der Online-Repositorien und der freien Webpublikationen übertragen werden kann. Hierbei sind zwei wichtige Punkte zu beachten, die, wie später in der Diskussion der Ergebnisse noch dargelegt werden wird, prinzipielle Unterschiede zwischen klassischen Zeitschriften und Dokumentenservern widerspiegeln. Zum einen der fachliche Schwerpunkt: Während man bei Zeitschriften in der Regel von einer fachlichen Spezifikation ausgehen kann, ist dies bei Dokumentenservern nur eingeschränkt möglich. Zwar gibt es auch hier durchaus das Prinzip der Konzentration auf einen fachlichen Schwerpunkt. Physikern fällt hier in der Regel der Preprintserver Arxiv ein, der, wenn er auch mittlerweile alle Bereiche der Physik und Mathematik abdeckt, seinen Schwerpunkt nach wie vor im Bereich der Hochenergie Teilchenphysik hat. Für Doku-

mentenserver, die von Universitäten bzw. Universitätsbibliotheken gepflegt werden, kann man von einer solchen fachlichen Fokussierung jedoch generell nicht ausgehen. Hier können in der Regel Wissenschaftler aller Fakultäten und Disziplinen ihre Arbeiten bereitstellen. Damit ist es zunächst fraglich, ob sich das erwartete, durch fachliche Spezifikation motivierte Potenzgesetzverhalten der Verteilung von Dokumenten auf Servern bestätigen lässt. Der zweite wichtige Unterschied zu Zeitschriften sind mögliche Dubletten. Während ein Zeitschriftenartikel per Definition eine eigenständige und unabhängige Einheit ist, die an einer klar definierten Stelle in die Verteilungsfunktion (und damit in das BLS) eingeht, erfolgt die Suche bei OAIster z.T. über nachgeordnete Instanzen die ihrerseits selbst den Inhalt anderer Repositorien enthalten können – sei es nun, dass sie deren Inhalt spiegeln oder selbst als Metasuchmaschine fungieren. Dadurch sind echte Dubletten möglich in der Gestalt, dass einige große Instanzen, die von OAIster abgefragt werden, ganze Sammlungen von Artikeln enthalten, die von OAIster noch einmal eigenständig abgefragt werden. Dies kann zur Folge haben, dass das BLS ebenfalls substantiell verfälscht wird.

Die Entnahme der Stichproben aus dem OAIster Datenpool gestaltete sich sehr viel einfacher als beim Web of Science. OAIster liefert selbst eine grobe Statistik, wie viele Datensätze in den jeweils unterschiedlichen Repositorien gefunden wurden. Diese wurde für die hier durchgeführten Analysen genutzt, so dass auf einen vollständigen Datensatz der Trefferliste verzichtet werden konnte. Somit konnte bei OAIster auch auf die zeitliche Beschränkung der Suche verzichtet werden. Dies stellt jedoch keinerlei Problem im Hinblick der Vergleichbarkeit zweier Stichproben zu einem Suchwort aus dem Web of Science und OAIster dar. Die Trefferlisten und damit die Verteilungsfunktionen sind ohnehin nur rein qualitativ vergleichbar. Wenn überhaupt, so würde ein detaillierter quantitativer Vergleich nur innerhalb der beiden Datenbanken sinnvoll sein.

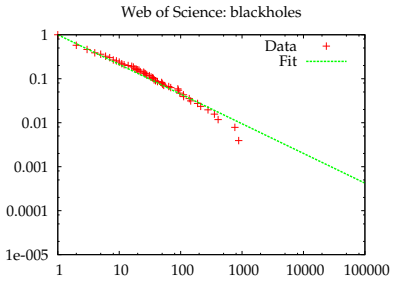
3.2 Darstellung und Auswertung

Im Folgenden werden die kumulierten Verteilungsfunktionen der gewonnenen Stichproben entsprechend dem im Kapitel 2 dargelegten Formalismus dargestellt. Hierbei werden jeweils die beiden Stichproben zu ein und demselben Suchwort gegenübergestellt. Aus Gründen der Übersicht erhält jede Stichprobe einen eigenen Graphen. Auf der x-Achse sind dabei jeweils die Artikelzahlen aufgetragen und auf der y-Achse die Anzahl $P(X \geq x)$. Diese Zahl entspricht in normierter Darstellung der Wahrscheinlichkeit, dass eine Zeitschrift mehr oder gleich x Artikel zu dem jeweiligen Thema hat. Da jede Kurve sowohl Wahrscheinlichkeitsverteilung als auch kumulierte Verteilungsfunktion der konkreten Stichprobe darstellt, liefert $y(x)$ natürlich auch den jeweiligen Bruchteil der Zeitschriften, die mehr als oder gleich x Artikel haben. Gezeigt werden jeweils die diskreten Punkte der kumulierten Verteilungsfunktion sowie eine angefittete Potenzfunktion. Der Exponent der gefitteten Funktion ist in der jeweiligen Bildunterschrift aufgeführt. Tabelle 3.1 fasst diese Angaben noch einmal zusammen und nennt zusätzlich noch den jeweiligen Stichprobenumfang. Zum Fit sei hier angemerkt, dass

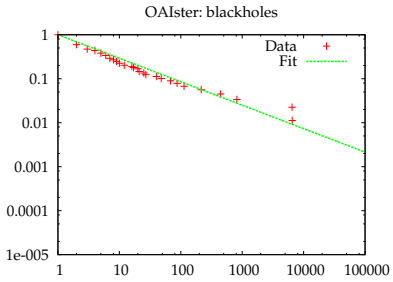
Suchbegriff	WoS			OAIster		
	Stichprobe	Abb. Nr.	Steigung	Stichprobe	Abb. Nr.	Steigung
blackholes	5612	3.1 a	-0.67	15178	3.1 b	-0.53
convection	3202	3.2 a	-1.12	30551	3.2 b	-0.60
education	15196	3.3 a	-1.20	145847	3.3 b	-0.31
empirical	12506	3.4 a	-1.45	51276	3.4 b	-0.49
hazard	7181	3.5 a	-1.23	6089	3.5 b	-0.68
human evolution	927	3.6 a	-1.28	271	3.6 b	-0.90
introduction	16259	3.7 a	-1.40	265635	3.7 b	-0.49
investigation	26042	3.8 a	-1.29	67273	3.8 b	-0.45
laser	24433	3.9 a	-0.92	36031	3.9 b	-0.49
music	14722	3.10 a	-0.97	95249	3.10 b	-0.46
optics	3031	3.11 a	-1.12	23595	3.11 b	-0.54
phase transition	25772	3.12 a	-0.76	21591	3.12 b	-0.58
philosophy	12910	3.13 a	-1.15	13946	3.13 b	-0.59
power law	10800	3.14 a	-1.00	9142	3.14 b	-0.63
revolution	10443	3.15 a	-1.34	10295	3.15 b	-0.64
self organization	3942	3.16 a	-1.23	17	3.16 b	-1.02
semiconductor	7474	3.17 a	-0.87	13379	3.17 b	-0.53
solid state physics	252	3.18 a	-2.02	632	3.18 b	-072
string theory	3775	3.19 a	-0.64	11808	3.19 b	-0.48
theoretical physics	206	3.20 a	-1.83	5873	3.20 b	-0.53

Tab. 3.1: Tabellarische Darstellung der untersuchten Stichproben

die Prozedur aufgrund der enormen Bandbreite der Rohdaten (in der Regel immerhin mehrere 10er Potenzen) nicht für die Rohdaten selbst durchgeführt wurde, sondern für die logarithmisierten Verteilungsfunktionen. Anschaulich gesprochen wurde also nicht versucht, eine Potenzfunktion anzupassen, sondern nach Umskalierung eine Gerade, die anschließend zurück transformiert wurde. Dieses Verfahren ist aus mathematischer Sicht durchaus kritikwürdig, gibt es doch weit bessere – dafür jedoch auch weit aufwendigere – Verfahren, um derartige Probleme zu lösen. Da wir uns im Weiteren jedoch weitgehend auf eine eher qualitative Diskussion und Betrachtung der Ergebnisse beschränken wollen, und insbesondere keine Messfehler zu berücksichtigen sind, kann dieses mathematisch eher laxe Verfahren durchaus als statthaft angesehen werden.

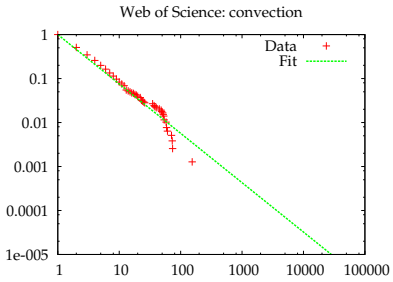


(a) WoS; $\alpha = -0.67$; $N = 5612$

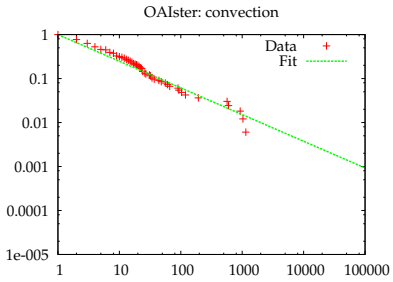


(b) OAIster; $\alpha = -0.53$; $N = 15178$

Abb. 3.1: Suchbegriff: blackholes

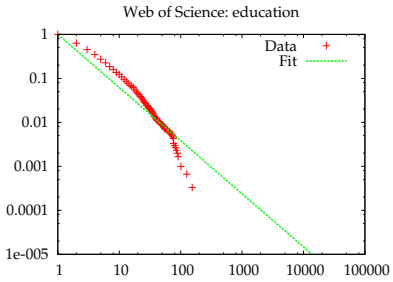


(a) WoS; $\alpha = -1.12$; $N = 3202$

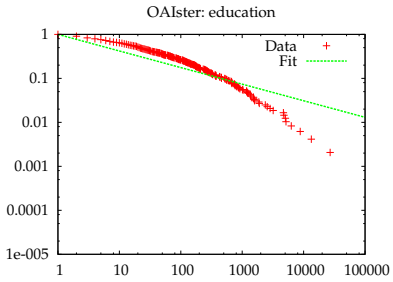


(b) OAIster; $\alpha = -0.60$; $N = 30551$

Abb. 3.2: Suchbegriff: convection

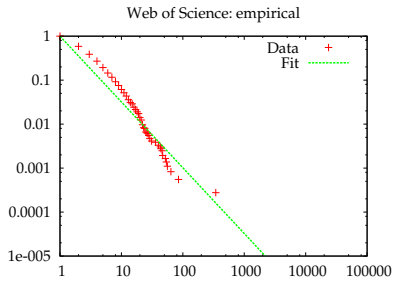


(a) WoS; $\alpha = -1.20$; $N = 15196$

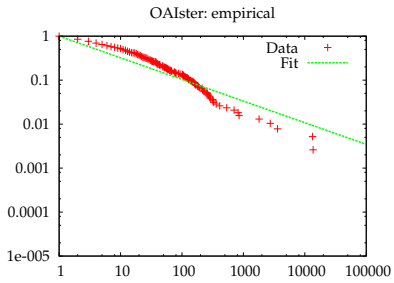


(b) OAIster; $\alpha = -0.31$; $N = 145847$

Abb. 3.3: Suchbegriff: education

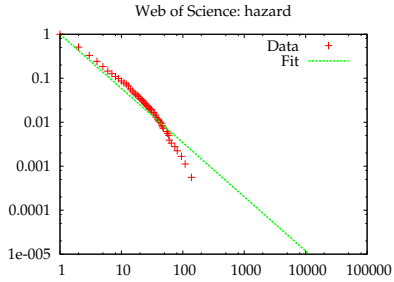


(a) WoS; $\alpha = -1.45$; $N = 12506$

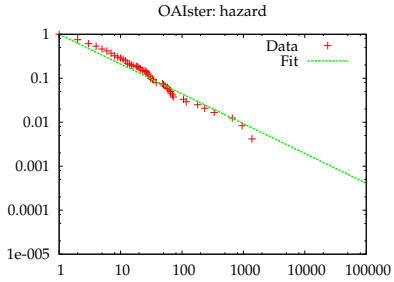


(b) OAIster; $\alpha = -0.49$; $N = 51276$

Abb. 3.4: Suchbegriff: empirical

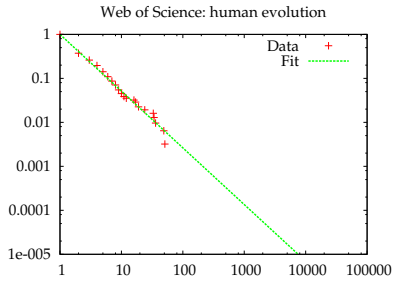


(a) WoS; $\alpha = -1.23$; $N = 7181$

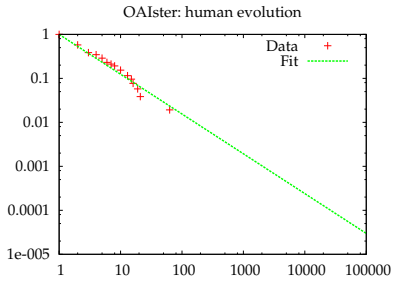


(b) OAIster; $\alpha = -0.68$; $N = 6089$

Abb. 3.5: Suchbegriff: hazard

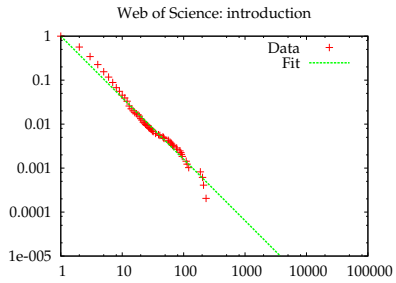


(a) WoS; $\alpha = -1.28$; $N = 927$

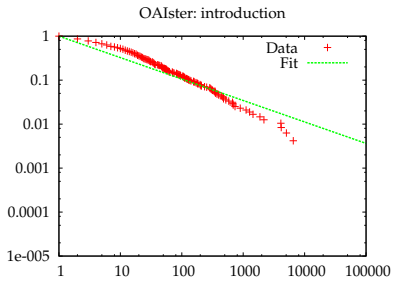


(b) OAIster; $\alpha = -0.90$; $N = 271$

Abb. 3.6: Suchbegriff: human evolution

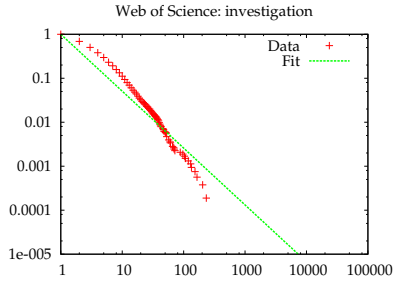


(a) WoS; $\alpha = -1.40$; $N = 16259$

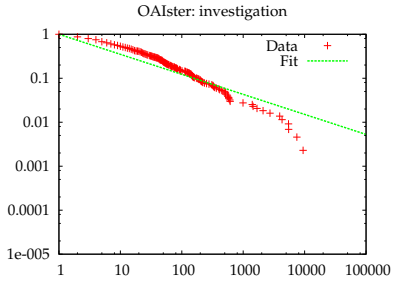


(b) OAIster; $\alpha = -0.49$; $N = 265635$

Abb. 3.7: Suchbegriff: introduction

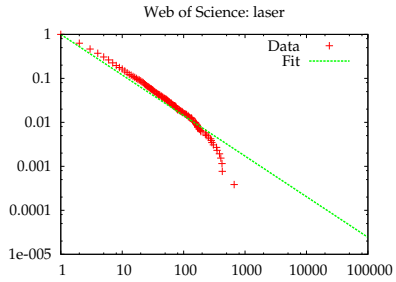


(a) WoS; $\alpha = -1.29$; $N = 26042$

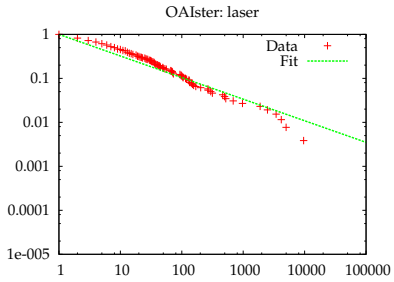


(b) OAIster; $\alpha = -0.45$; $N = 67273$

Abb. 3.8: Suchbegriff: investigation

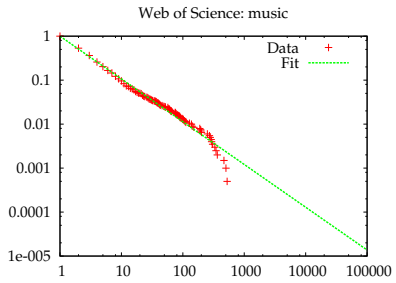


(a) WoS; $\alpha = -0.92$; $N = 24433$

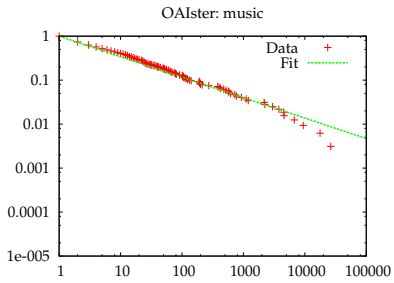


(b) OAIster; $\alpha = -0.49$; $N = 36031$

Abb. 3.9: Suchbegriff: laser

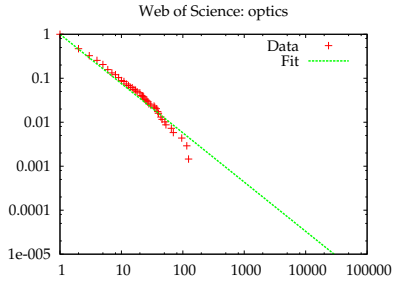


(a) WoS; $\alpha = -0.97$; $N = 14722$

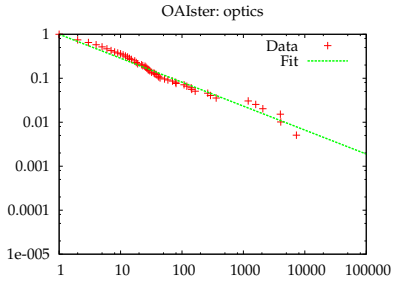


(b) OAIster; $\alpha = -0.46$; $N = 95249$

Abb. 3.10: Suchbegriff: music

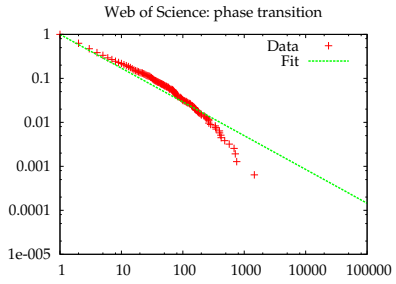


(a) WoS; $\alpha = -1.12$; $N = 3031$

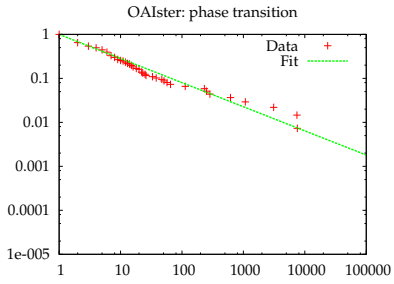


(b) OAIster; $\alpha = -0.54$; $N = 23595$

Abb. 3.11: Suchbegriff: optics

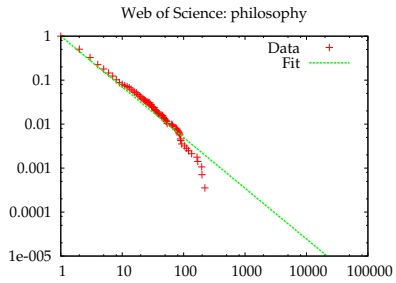


(a) WoS; $\alpha = -0.76$; $N = 25772$

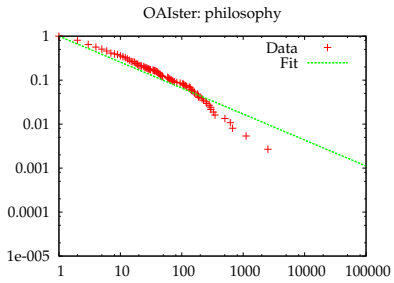


(b) OAIster; $\alpha = -0.58$; $N = 21591$

Abb. 3.12: Suchbegriff: phase transition

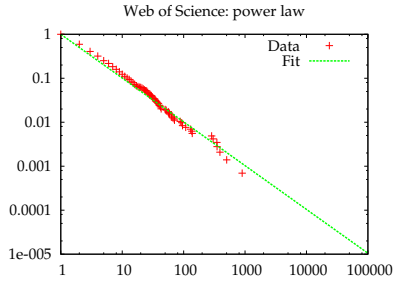


(a) WoS; $\alpha = -1.15$; $N = 12910$

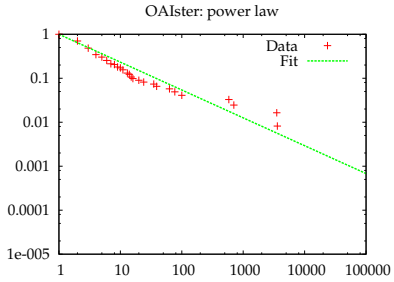


(b) OAIster; $\alpha = -0.59$; $N = 13946$

Abb. 3.13: Suchbegriff: philosophy

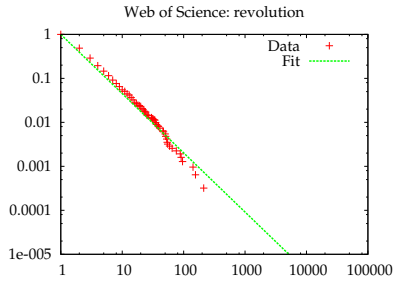


(a) WoS; $\alpha = -1.00$; $N = 10800$

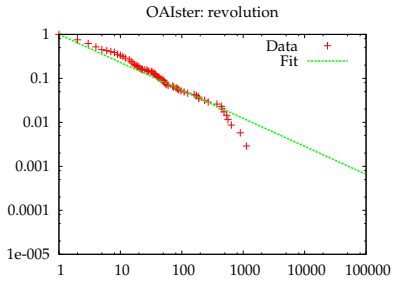


(b) OAIster; $\alpha = -0.63$; $N = 9142$

Abb. 3.14: Suchbegriff: power law

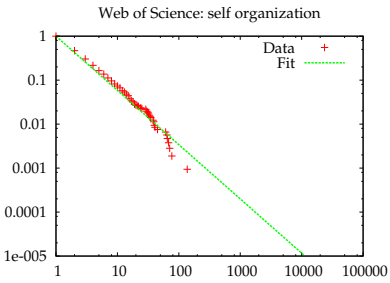


(a) WoS; $\alpha = -1.34$; $N = 10443$

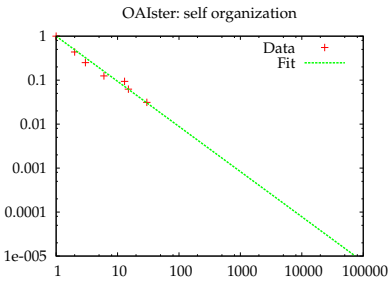


(b) OAIster; $\alpha = -0.64$; $N = 10295$

Abb. 3.15: Suchbegriff: revolution

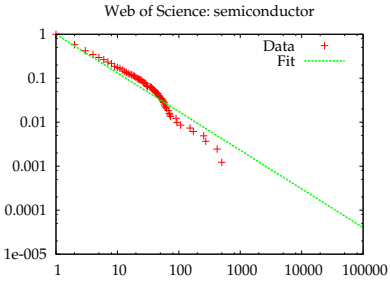


(a) WoS; $\alpha = -1.23$; $N = 3942$

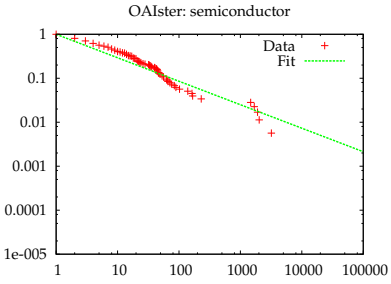


(b) OAIster; $\alpha = -1.02$; $N = 17$

Abb. 3.16: Suchbegriff: self organisation

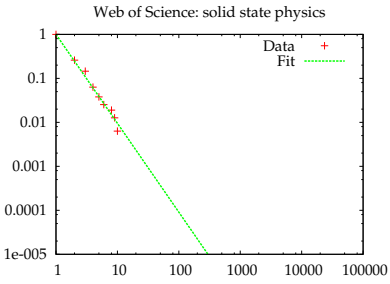


(a) WoS; $\alpha = -0.87$; $N = 7474$

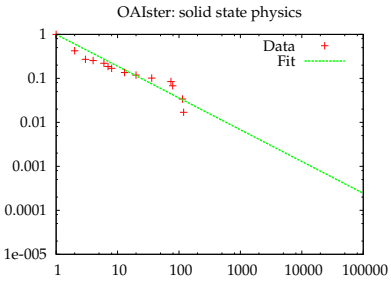


(b) OAIster; $\alpha = -0.53$; $N = 13379$

Abb. 3.17: Suchbegriff: semiconductor

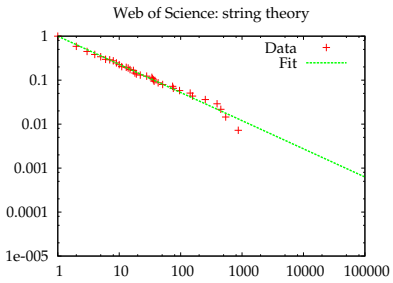


(a) WoS; $\alpha = -2.02$; $N = 252$

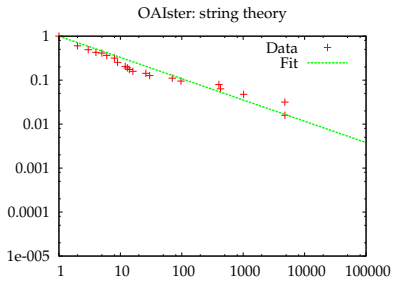


(b) OAIster; $\alpha = -0.72$; $N = 632$

Abb. 3.18: Suchbegriff: solid state physics

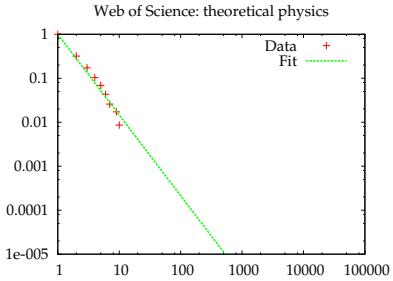


(a) WoS; $\alpha = -0.64$; $N = 3775$

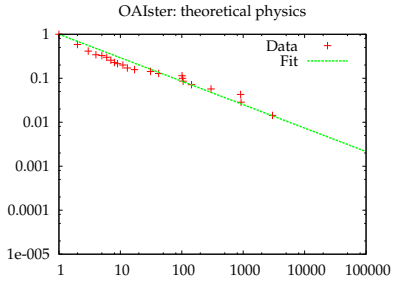


(b) OAIster; $\alpha = -0.48$; $N = 11808$

Abb. 3.19: Suchbegriff: string theory



(a) WoS; $\alpha = -1.83$; $N = 206$



(b) OAIster; $\alpha = -0.53$; $N = 5837$

Abb. 3.20: Suchbegriff: theoretical physics

3.3 Ergebnisse und Diskussion

Die Hauptergebnisse der Sichtung der gewonnenen Stichproben lassen sich wie folgt zusammenfassen:
1) Das BLS zur Beschreibung der Verteilung wissenschaftlicher Publikationen auf Zeitschriften bzw. Dokumentenservern kann bedingt als erste Näherung bestätigt werden.
2) Die vorliegende Datenlage gibt keinerlei Hinweis auf einen systematischen Unterschied zwischen den Verteilungen zu eher fachspezifischen und eher nicht fachspezifischen Suchbegriffen.
3) Die jeweiligen Daten von Web of Science und OAIster zu ein und demselben Suchwort weisen keine signifikante Korrelation auf.
4) Es gibt einen prinzipiellen Unterschied zwischen den beiden Datenbanken, der bei OAIster zu systematisch flacheren Verteilungsfunktionen führt.
5) Es existieren neben dem teilweise nur rudimentär vorhandenen Potenzgesetzverhalten der Verteilungsfunktionen z.T. erhebliche Abweichungen von diesem. Hier sind im Wesentlichen drei sich wiederholende Charakteristika zu finden.
Im Folgenden soll auf die einzelnen Punkte ausführlicher eingegangen werden:
1) Das BLS, ursprünglich in einer eher qualitativen Form beschrieben, hat niemals den Anspruch erhoben, eine Art exaktes Naturgesetz zu sein. Unter dieser Prämisse können die erhobenen Stichproben durchaus als Bestätigung des grundlegenden Potenzgesetzverhaltens der Verteilungsfunktion gewertet werden. Es sei ausdrücklich

betont, dass, obgleich die doppelt logarithmische Darstellung der Verteilungsfunktionen durchaus ihren in Kapitel 2 beschriebenen Sinn hat, die Darstellung auch dazu führen kann, etwaige Abweichungen vom echten Potenzgesetz zu unterschätzen. Dass in dieser Arbeit dennoch jeweils eine Potenzfunktion an die Stichprobenergebnisse angefittet wurde – unter Inkaufnahme teilweise erheblicher Fehlermargen – unterstreicht nur den qualitativen Charakter, der hier bei der Beschreibung und Überprüfung des BLS gelten soll. Reduziert und vereinfacht man die Aussage des BLS auf die Existenz einiger Kernzeitschriften und die vieler Randzeitschriften, deren Verteilung in etwa einem Potenzgesetz folgt, so kann diese Aussage durch die gewonnenen Stichproben unterstützt werden. Dies soll natürlich nicht die Frage verdrängen, warum und inwieweit ein abweichendes Verhalten zwischen den Stichproben und dem theoretischen Verhalten zu finden ist.
2) Ein sicherlich überraschendes, wenn auch aufgrund der geringen Anzahl der Stichproben nicht sehr belastbares Ergebnis ist, dass sich kein systematischer Unterschied zwischen den Verteilungen zu den eher fachspezifischen Suchbegriffen auf der einen Seite und den eher nicht fachspezifischen Suchbegriffen auf der anderen Seite feststellen lässt. Die Anzahl von 20 Stichproben je Quelle und je 3 unspezifischen und je 17 fachspezifischen Suchbegriffen ist jedoch bei weitem zu gering, um hier eine aussagekräftige Statistik zu erwarten. Als unspezifische Begriffe sollen hier empirical, introduction und investigation gelten. Gleiches gilt für den Unterschied zwischen den verschiedenen fachspezifischen Suchbegriffen. Wünschenswert wäre hier sicherlich, einen umfassenden Scan – etwa mit der DDC als Grundlage – der beiden Datenbanken durchzuführen. Vorstellbar wäre es z.B., ab der dritten Ebene der DDC (1000 Suchbegriffe) – oder einer vergleichbaren Systematik – nach systematischen und fachspezifischen Unterschieden zwischen den Verteilungen der verschiedenen Disziplinen zu suchen. Ein solch umfangreiches Programm kann im Rahmen dieser Masterarbeit aber nicht durchgeführt werden.
3) Im Verhältnis WoS zu OAIster lässt sich ebenfalls keine belastbare Korrelation der Daten finden. Eine Untersuchung, ob Verteilungen, die im WoS eine bestimmte Charakteristik aufweisen, die gleiche Charakteristik mit einer signifikant höheren Wahrscheinlichkeit auch in OAIster aufweisen, würde gleichfalls (siehe Punkt 2) eine sehr viel größere Anzahl an Stichproben erfordern. Nichtsdestotrotz wäre eine derartige Untersuchung absolut notwendig, um die Übertragbarkeit und Allgemeingültigkeit gefundener Charakteristika abschätzen zu können.
4) Über den prinzipiellen Mechanismus, der im Verhältnis WoS und OAIster zu generell sehr viel flacheren Verteilungsfunktionen bei OAIster führt, kann bei der vorhandenen Datenlage nur spekuliert werden. Als wahrscheinliche und naheliegende Ursache hierfür muss das Faktum angesehen werden, dass bei OAIster grundsätzlich die Möglichkeit von Dubletten gegeben ist. Dubletten von Artikeln sind in OAIster dadurch prinzipiell gegeben, dass in dieser Datenquelle dezentral gepflegte und unabhängig verwaltete Dokumentenserver abgefragt werden. Dies hat zur Folge, dass ein und derselbe Artikel von Autoren auf verschiedenen Servern abgelegt werden kann und damit zweifach in OAIster auftaucht. Zudem können die Inhalte von Servern ganz oder teilweise Teilmengen anderer Server sein und damit von OAIster doppelt abgefragt und gelistet werden. Das bedeutet: Ist der Dokumentenserver A Teilmenge von Server B und gehen beide in die Verteilungsfunktion ein, so entspricht dies einer ähnli-

chen Integration, wie sie beim Übergang von der Verteilungsfunktion zur kumulierten Verteilungsfunktion (siehe Kapitel 2) erfolgt. Das Resultat ist eine systematisch flachere Verteilung bei OAIster.
5) Das interessanteste Ergebnis der vorliegenden Untersuchung ist sicherlich, dass es sich wiederholende und charakteristische Unterschiede zwischen den dargestellten Stichproben und einem echten Potenzgesetz gibt. Hier sind in erster Linie drei Charakteristika zu nennen: Zunächst eine sicherlich rein statistisch zu begründende unregelmäßige Streuung der Verteilungen um die angefitteten Potenzfunktionen, die vor allem am rechten Rand der Verteilungsfunktion, d.h. bei großen Artikelzahlen, auftritt. Hierzu kann wenig mehr gesagt werden, als dass diese Punkte durchweg eine sehr geringe Anzahl an Zeitschriften repräsentieren. Namentlich sind dies die Kernzeitschriften für das jeweilige Thema. Eine Streuung realer Werte um eine idealisierte Funktion ist hier unbedingt zu erwarten und alles andere als erstaunlich. Als zweites Charakteristikum ist eine klare und systematische Abweichung vom Potenzgesetz zu nennen. Diese findet sich in den Verteilungen zu diversen Stichproben. Etwa: education (WoS/OAIster), empirical (WoS/OAIster), hazard (WoS), investigation (WoS/OAIster) und phase transition (WoS). Hier zeigt sich weniger ein linearer Abfall der Verteilung im log-log-Diagramm als vielmehr eine leicht konkave Kurve mit vom Betrag her wachsender negativer Steigung. Wie später in Kapitel 4 noch gezeigt werden wird, deutet dies, ohne dass hier eine genaue statistische Analyse erstellt werden soll, darauf hin, dass es sich hier tendenziell eher um Exponentialfunktionen denn um Potenzfunktionen handelt. Gleich um welchen Funktionstyp es sich auch handelt, die grundlegende Aussage des BLS bleibt in jedem Falle erhalten. Ebenfalls in Kapitel 4 wird gezeigt werden, dass eine leichte Modifizierung des zur Beschreibung des BLS herangezogenen Modells auf einfache Weise zu Verteilungen mit beiden Funktionstypen führen kann. Der dritte wichtige Punkt bei der Diskussion der Stichproben ist das fast überall zu findende verstärkte Abfallen der Daten am rechten Rand der Verteilung im Vergleich zur reinen Potenzfunktion. Besonders ausgeprägt ist dies Verhalten bei den Stichproben zu laser (WoS/OAIster), music (WoS/OAIster) und investigation (OAIster) zu finden. Teilweise wird dieses Verhalten von dem als zweites Charakteristikum genannte Abweichen von der Potenzfunktion verstärkt und überlagert. Auch für diesen Effekt soll im Kapitel 4 im Rahmen des dort definierten numerischen Modells ein Erklärungsvorschlag formuliert werden.

Kapitel 4

Theorie, Modellierung und Simulation

Der Prozess, der zum BLS führt, d.h. die Verteilung von Publikationen auf Zeitschriften bzw. Dokumentenservern, soll nun numerisch modelliert werden. Hierzu ziehen wir das Yule-Simon-Modell heran [*Simon*, 1955], das bereits erfolgreich im Bereich der Bibliometrie [*Chen*, 1995]; [*Chen*, 1986] und diversen anderen Bereichen zur Beschreibung von ähnlichen Prozessen benutzt wurde [*Bornholdt*, 2001]; [*Pennock*, 2002]; [*Levene*, 2006]. Bevor der mathematische Formalismus des Modells eingeführt wird, soll hier kurz eine anschauliche nicht mathematische Beschreibung der Prozedur erfolgen. Das Modell geht von der Annahme aus, dass ein Wissenschaftler einen Artikel vorzugsweise in einer Zeitschrift veröffentlicht, die bereits viele Artikel zum gleichen Thema herausgebracht hat. Um aus dieser Aussage ein programmierbares Modell zu formulieren, muss ihre Kernaussage präzise herausgearbeitet werden: „Vorzugsweise" bedeutet in diesem Zusammenhang „mit größerer Wahrscheinlichkeit". Dies beinhaltet, damit das Modell auch bisher nicht besetzte Zeitschriften berücksichtigen kann, dass die Veröffentlichung immer auch mit einer von Null verschiedenen Wahrscheinlichkeit in Zeitschriften erfolgen kann, die weniger oder bisher keine Artikel zum Thema enthalten. Weiter geht das Modell davon aus, dass der Wissenschaftler, der die Entscheidung trifft, welche Zeitschrift er wählt, einen vollständigen Überblick über die Verteilung aller bisher zu seinem Thema erschienenen Artikel besitzt. Es ist offensichtlich, dass es sich hierbei um starke Vereinfachungen des realen Publikationsprozesses handelt, der mit all seinen detaillierten Komponenten und zu treffenden Entscheidungen sicherlich nicht durch ein Computerprogramm abgebildet werden kann. Es wird sich zeigen, inwieweit dieses vereinfachte Modell in der Lage ist, die empirischen Daten aus Kapitel 3 zu reproduzieren. Das Programm läuft nun im Einzelnen wie folgt ab: Ein Wissenschaftler trifft die Entscheidung, einen Artikel zu veröffentlichen. Er verschafft sich einen vollständigen Überblick über die schon erschienenen Artikel zu seinem Thema. Dies bedeutet, er weiß, wie viele Artikel zu seinem Thema von den einzelnen Zeitschriften schon veröffentlicht wurden. Er unterscheidet die Menge an Zeitschriften, die schon Artikel zum Thema enthalten (A) und die, die noch keine enthalten (B). Mit einer im Allgemeinen sehr geringen Wahrscheinlichkeit q entscheidet er sich, in einer Zeitschrift der Menge B zu veröffentlichen. Diese Zeitschrift wechselt damit automatisch aus Menge B in die Menge A. Entscheidet er sich mit der Wahrscheinlichkeit $p = 1 - q$, in der Menge A zu veröffentlichen, so muss er festlegen, welche dieser Zeitschriften er auswählt. Dies erfolgt wiederum auf der Basis von Wahrscheinlichkeiten. Die Wahr-

scheinlichkeit, in einer bestimmten Zeitschrift aus Gruppe B zu veröffentlichen, ist dabei im Basismodell proportional zur Anzahl der in dieser Zeitschrift (zum jeweiligen Thema) bereits veröffentlichten Artikel. Für diesen Ansatz hat sich in der Literatur der Begriff des „preferential attachments“ herausgebildet. Eine Normierung erfolgt dadurch, dass nach der Entscheidung für Gruppe B die Wahrscheinlichkeit, überhaupt in einer dieser Zeitschriften zu publizieren, sicher (also gleich 1) ist. Dieser Ablauf wird dann N-mal iteriert, wobei N die Anzahl der zu veröffentlichenden Artikel bezeichnet. Es ist bemerkenswert, dass der reine Yule-Simon-Prozess analytisch lösbar ist. Bevor das numerische Programm vorgestellt wird, soll daher die analytische Lösung dargelegt werden. Obgleich der dafür notwendige mathematische Apparat im Rahmen dieser Arbeit nicht in aller Ausführlichkeit eingeführt werden kann, soll dennoch im nächsten Abschnitt auf grundlegende Definitionen und Gleichungen eingegangen werden.

4.1 Mastergleichung

Der beschriebene Prozess der Verteilung von Artikeln auf eine Menge von Zeitschriften ist von seiner Natur her ein Markov-Prozess, oder genauer, da es sich um einen unstetigen, diskreten Prozess handelt, eine Markov-Kette. Bei der Untersuchung ist nun das Ziel, die Wahrscheinlichkeitsverteilung für die Verteilung von Artikeln auf Zeitschriften anzugeben. Da, wie gerade beschrieben, der Prozess keinerlei Gedächtnis hat, erfolgt die Auswahl einer Zeitschrift durch einen Wissenschaftler unabhängig von der Geschichte des Prozesses. Es geht lediglich die absolute Zahl der Artikel pro Zeitschrift in die Entscheidung ein. Daher bietet sich für eine Beschreibung die Mastergleichung an. Diese soll in der gebotenen Kürze hergeleitet werden.

Es sei $p(A)$ die Wahrscheinlichkeit, dass das Ereignis A eintreffe. $p(A|B)$ wird bedingte Wahrscheinlichkeit genannt und bezeichnet die Wahrscheinlichkeit des Ereignisses A unter der Bedingung, dass das Ereignis B bereits eingetroffen ist.

Mit dieser Schreibweise kann nun die grundlegende Markov-Eigenschaft eines Prozesses definiert werden. Ganz allgemein kann nun die Wahrscheinlichkeit für ein Ereignis (A und B und C und D ...) auf verschiedene Weisen faktorisiert werden. Als Referenz mag hier auf ein beliebiges Textbuch der statistischen Physik verwiesen sein. Als Beispiel soll nur das bewährte Standardwerk [*Gardiner*, 2004] genannt werden.

$$\begin{aligned} p(A,B,C,D,...) &= p(A)p(B|A)p(C|A,B)p(D|A,B,C)... \\ &= p(B)p(A|B)p(C|A,B)p(D|A,B,C)... \\ &= p(C)p(A|C)p(B|A,C)p(D|A,B,C)... \end{aligned} \tag{4.1}$$

Der Prozess ist dann markovsch, wenn es eine ausgezeichnete Anordnung gibt (z.B. zeitlich bedingt), für die sich die Darstellung vereinfacht, so dass gilt:

$$p(A,B,C,D,...) = p(A)p(B|A)p(C|B)p(D|C) \tag{4.2}$$

Gilt nun für die zeitliche Abfolge der Ereignisse $t(A) < t(B) < t(C) < t(D) < ...$, so folgt, dass die Wahrscheinlichkeit für den nächsten Zustand nur vom jeweils vorausgehenden abhängt. Beschreibe nun das Paar (x_i, t_i) das Ereignis x_i zum Zeitpunkt t_i. Damit erhalten wir dann nach gewissen Umformungen die Smoluchowski-Chapman-Kolmogorov-Gleichung:

$$p(x_3, t_3|x_1, t_1) = \int dx_2 p(x_3, t_3|x_2, t_2) p(x_2, t_2|x_1, t_1) \tag{4.3}$$

Per Definition ist die Übergangsrate von $\acute{x}$ nach x gegeben durch:

$$W_{t,}(x|\acute{x}) = \frac{\partial p(x, t|\acute{x}, \acute{t})}{\partial t} \tag{4.4}$$

Nach weiteren Umformungen, für die wieder auf ein beliebiges Standardwerk der statistischen Physik verwiesen sei, erhält man die Mastergleichung:

$$\frac{\partial p(x, t|x_0, t_0)}{\partial p} = \sum_{\acute{x}} (W_t(x|\acute{x}) p(\acute{x}, t|x_0, t_0) - (W_t(\acute{x}|x) p(x, t|x_0, t_0)) \tag{4.5}$$

Es sei noch einmal ausdrücklich darauf hingewiesen, dass dies selbstverständlich keine vollständige Herleitung der Mastergleichung ist, sondern allenfalls eine grobe Skizze derselben. Es soll jedoch noch etwas genauer auf die Bedeutung und Interpretation dieser für die Statistische Physik so überaus wichtigen Gleichung eingegangen werden. Die Mastergleichung ist ebenso wie die Smoluchowski-Chapman-Kolmogorov-Gleichung eine Konsistenzbedingung für die enthaltenen bedingten Wahrscheinlichkeiten $p(x|\acute{x})$. Die Gültigkeit dieser Konsistenzbedingung ist allerdings an die Voraussetzung geknüpft, dass der beschriebene Prozess die Markov-Eigenschaft besitzt. Diese besagt, um dies noch einmal hervorzuheben, dass bei einer zeitlichen Folge von Ereignissen ein Ereignis lediglich vom jeweils vorherigen Zustand abhängt. Somit stellt die Markov-Eigenschaft eine minimale Version der „Erinnerung“ eines Prozesses an seine Vergangenheit dar. Die Markov-Annahme ist eine Idealisierung für eine Reihe von physikalischen Prozessen, deren Berechtigung und Gültigkeit natürlich jeweils genau zu prüfen ist. Als prominentestes Beispiel für einen Markov-Prozess gilt sicherlich die Brownsche Bewegung, d.h. die Bewegung eines großen (makroskopisch beobachtbaren) Teilchens in einer Flüssigkeit. Die große Bedeutung der Mastergleichung liegt darin, dass aus vorgegebenen Übergangswahrscheinlichkeiten die bedingten Wahrscheinlichkeiten berechnet werden können. Dies geschieht in der Regel durch die Lösung eines Systems gekoppelter Differentialgleichungen. Das vorliegende Problem kann jedoch mit Hilfe einfacherer Algebra und Kombinatorik gelöst werden. Die benötigten Übergangswahrscheinlichkeiten werden dabei direkt dem Modell, d.h. den zugrundeliegenden Annahmen über das Verhalten von Wissenschaftlern beim Publikationsprozess, entnommen. Die resultierenden bedingten Wahrscheinlichkeiten entsprechen der Verteilungsfunktion bzw. der Wahrscheinlichkeitsdichte für die Verteilung von Artikeln auf Zeitschriften. Die vorgestellte Mastergleichung stellt nur ein,

noch dazu relativ simples Werkzeug dar, das sich in der Statistischen Physik zur Lösung stochastischer Zufallsprozesse bewährt hat. Im Ausblick am Ende dieser Arbeit soll noch eingehender auf die mannigfaltigen Möglichkeiten, die die Statistische Physik für bibliometrische Fragestellungen bereithält, hingewiesen werden. An dieser Stelle sei nur angedeutet, dass vermöge der Fokker-Plank-Gleichung und des Lemma von Ito eine enge Verwandtschaft zwischen der Mastergleichung und dem Kalkül stochastischer Differentialgleichungen besteht. Damit erschließt sich für die bibliometrische Forschung ein sehr viel mächtigerer mathematischer Apparat als es das hier vorgestellte Modell vermuten lässt.

4.1.1 Analytische Lösung

Um den zu Beginn dieses Kapitels beschriebenen Yule-Simon-Prozess nun mit Hilfe der Mastergleichung zu lösen, muss der schon anschaulich beschriebene Ablauf formalisiert werden. Die vorgestellte Darstellung richtet sich erneut nach [*Newman*, 2005].
Hierzu sei Folgendes angenommen:
Artikel können entstehen, jedoch nicht wieder verschwinden. Sie müssen einer und nur einer Zeitschrift zugeordnet werden (von der auf Dauer mühseligen Doppelnennung von Zeitschrift und Dokumentenserver soll hier bis auf Weiteres abgesehen werden). Artikel werden mit der Wahrscheinlichkeit q bei Zeitschriften veröffentlicht, die noch keine Artikel zum Thema haben.
Sei $p_{k,n}$ der Anteil an Zeitschriften mit genau k Artikeln zum relevanten Thema. n bezeichne dabei die Gesamtzahl aller Zeitschriften. Die Anzahl der Zeitschriften mit k Artikeln ist damit $np_{k,n}$.
Gesucht ist nun die Wahrscheinlichkeit, dass der nächste zu veröffentlichende Artikel einer bestimmten Zeitschrift zugeführt wird. Diese Wahrscheinlichkeit wird hier als proportional zu den schon enthaltenen Artikeln angenommen. Sie beträgt damit k_i, bzw. normiert $k_i/(\sum_i k_i)$ wobei $\sum_i k_i$ lediglich die Gesamtzahl aller Artikel angibt. Diese ist wiederum gleich n/q. Sei nun $m = 1/q - 1$, so gilt $\sum_i k_i = n(m+1)$.
Zwischen der Erstbesetzung der n-ten Zeitschrift und der (n+1)ten werden im Mittel $1/q - 1 = m$ neue Artikel dem System zugeführt.
(Als kleines Zahlenbeispiel: Sei $q = 0.1$, so wird bei jedem zehnten Artikel eine neue Zeitschrift ausgewählt, die bisher noch keine Artikel zum Thema hatte. Bis dahin werden also $1/0.1 - 1 = 9$ Artikel auf die schon besetzten Zeitschriften verteilt.)
Die Wahrscheinlichkeit, für schon besetzte Zeitschriften einen neuen Artikel hinzuzubekommen, ist damit im Intervall zwischen zwei Erstbesetzungen einer Zeitschrift für die Zeitschrift i gleich $mk_i/(n(m+1))$. Somit folgt für die Gesamtzahl der zu erwartenden neuen Artikel für Zeitschriften mit bereits k Artikeln im gleichen Intervall:

$$\frac{mk}{n(m+1)} np_{k,n} = \frac{m}{m+1} kp_{k,n} \tag{4.6}$$

Die Anzahl der Zeitschriften mit k Artikeln f_k wird jedoch im selben Intervall um die gleiche Zahl reduziert, da sie ja nach Hinzufügung eines neuen Artikels nicht mehr länger zur Gruppe der Zeitschriften mit k, sondern zu der mit $k+1$ Artikeln gehören.

Zur gleichen Zeit wird sich jedoch die Zahl f_k wieder erhöhen, da schließlich Zeitschriften der Gruppe f_{k-1} ebenfalls neue Artikel erhalten.
Damit sind die Übergangswahrscheinlichkeiten für die Formulierung der Mastergleichung festgelegt. Sie lautet:

$$(n+1)p_{k,n+1} = np_{k,n} + \frac{m}{m+1}[(k-1)p_{k-1,n} - kp_{kpn}] \tag{4.7}$$

Die einzige Ausnahme davon gilt für $k = 1$:

$$(n+1)p_{1,n+1} = np_{1,n} + 1 - \frac{m}{m+1}p_{1,n} \tag{4.8}$$

Das eigentliche Interesse richtet sich jedoch auf die Verteilungsfunktion, die sich bei großen Artikelzahlen ergibt. Hierzu bilden wir den Limes $n \to \infty$ und nehmen an, dass die resultierende Verteilung unabhängig von n ist: $p_k = \lim_{n\to\infty} p_{n,k}$. Diese Näherung ist nach [*Schweitzer*, 1998] streng nur gültig, wenn $q << 1$. Damit folgt für $k = 1$:

$$p_1 = 1 - \frac{mp_1}{m+1} \tag{4.9}$$

oder

$$p_1 = \frac{m+1}{2m+1} \tag{4.10}$$

und für den allgemeinen Fall:

$$p_k = \frac{m}{m+1}[(k-1)p_{k-1} - kp_k] \tag{4.11}$$

oder

$$p_k = \frac{k-1}{k+1+1/m}p_{k-1} \tag{4.12}$$

Iteriert man diese Formel, ergibt sich:

$$\begin{aligned} p_k &= \frac{(k-1)(k-2)...1}{(k+1+1/m)(k+1/m)...(3+1/m)}p_1 \\ &= (1+1/m)\frac{(k-1)...1}{(k+1+1/m)...(2+1/m)} \end{aligned} \tag{4.13}$$

An dieser Stelle benutzen wir die bekannte Γ-Funktion:

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt \tag{4.14}$$

für die $\Gamma(a) = (a-1)\Gamma(a-1)$ und $\Gamma(1) = 1$ gilt. Somit erhalten wir:

$$p_k = (1+1/m)\frac{\Gamma(k)\Gamma(2+1/m)}{\Gamma(k+2+1/m)} \tag{4.15}$$

$$= (1+1/m)B(k, 2+1/m) \tag{4.16}$$

wobei $B(a,b)$ die Betafunktion bezeichnet, die mit $B(a,b) \sim a^{-b}$ angenähert werden kann. Der Exponent der Wahrscheinlichkeitsverteilungsfunktion ergibt sich damit zu:

$$\alpha = 2 + \frac{1}{m} \tag{4.17}$$

Bei der hier aus dem Yule-Simon-Modell hergeleiteten Betafunktion handelt es sich exakt um die gleiche Funktion, die schon in Kapitel 2 für die Verteilung einer diskreten Potenzgesetzverteilung gefunden wurde. Das Yule-Simon-Modell besitzt damit für große k als Lösung eine Potenzfunktion mit der Steigung $-\alpha = -(2+1/m)$. Da $m = 1/q-1$ gilt, folgt:

$$\alpha = 2 + \frac{1}{\frac{1}{q} - 1} \tag{4.18}$$

und damit für die Funktion:

$$f(x) = x^{-(2+1/(1/q-1))} \tag{4.19}$$

Hierbei ist q die Wahrscheinlichkeit für die Erstbesetzung einer Zeitschrift. Die kumulierte Verteilungsfunktion ergibt sich nun zu:

$$P[X \geq x] = x^{-1-1/(1/q-1)} \tag{4.20}$$

Es gilt: Bei wachsendem q wird der gesamte Exponent vom Betrag her größer, d.h. die Verteilungsfunktion fällt stärker ab. Da q aber auch ein Maß für die Interdisziplinarität von Themen ist (ist q groß, ist das entsprechende Fachgebiet als stark interdisziplinär aufzufassen), folgt konsistent mit der Aussage der Verteilungsfunktion: Eine steile Verteilung repräsentiert ein eher interdisziplinäres Thema (großes q), eine flachere Verteilung repräsentiert ein stark konzentriertes Fachgebiet (kleines q).

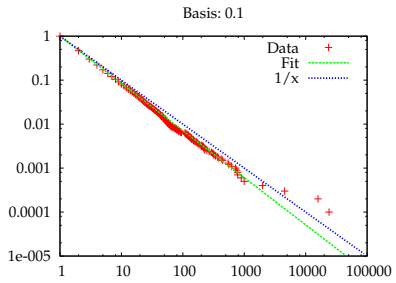
4.2 Konzeption der numerischen Modelle

Die Kodierung des vorgestellten stochastischen Prozesses erfolgte in der Programmiersprache C++. Es wurde für alle folgenden Simulationen jeweils die Verteilung von 20.000 Artikeln simuliert. Die Wahl dieser willkürlichen Zahl erfolgte im Hinblick auf die Stabilität der numerischen Lösung, einer Begrenzung der Rechenzeit und einer

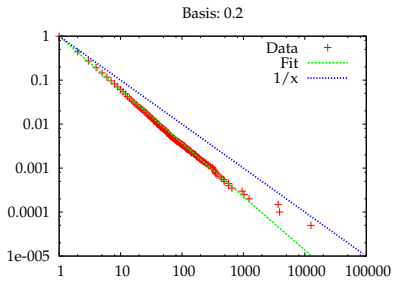
Vergleichbarkeit mit den Umfängen der gewonnenen empirischen Stichproben. Als Anfangsbedingung wurde je eine Zeitschrift mit einem Artikel besetzt. Die Basis-Version entspricht dem in Abschnitt 4.1.1 beschriebenen Yule-Simon-Prozess. Da dieser Prozess theoretisch zu einer Verteilung führt, die einem Potenzgesetz entspricht, wurden, um die in Kapitel 3 gefundenen und diskutierten Abweichungen durch das numerische Modell erklären zu können, zwei Erweiterungen zum Yule-Simon-Prozess definiert und programmiert. Die Einfachheit und hohe Abstraktion des Ansatzes führt dazu, dass nur eine sehr geringe Freiheit im Modell verbleibt. In der Basis-Version sowie in Erweiterung 1 (veränderte Wahrnehmung der Zeitschriften aufgrund der Gewichtung durch Wissenschaftler) steht nur ein freier Parameter zur Verfügung. Dieser entspricht der Wahrscheinlichkeit q, dass ein Artikel in einer Zeitschrift veröffentlicht wird, die noch keine Artikel zum Thema besitzt. In Erweiterung 2 kommt als zweiter freier Parameter die Ablehnwahrscheinlichkeit durch Zeitschriften hinzu. Im Folgenden soll zunächst der Programmablauf schematisch dargestellt werden. Der Programmablauf ist für alle drei Versionen analog und wird durch einen Quellcode beschrieben. Die Erweiterungen können jeweils zugeschaltet bzw. abgeschaltet werden.
Die Anfangskonfiguration ist jeweils dadurch gegeben, dass eine Zeitschrift mit einem Artikel besetzt ist. Die Verteilung der Artikel erfolgt dann in einer Schleife mit N Durchläufen, wobei N die Anzahl der simulierten Zeitschriftenartikel ist (hier also 20.000). Zunächst wird mit Hilfe eines Zufallszahlengenerators und der vorgegebenen Neubesetzungswahrscheinlichkeit q bestimmt, ob eine bisher nicht besetzte Zeitschrift mit einem Artikel besetzt wird. Wenn ja, so ist die Schleife beendet und wird erneut durchlaufen. Wenn nein, wird anhand der momentanen Besetzung der Zeitschriften ebenfalls mit Hilfe eines Zufallszahlengenerators bestimmt, welche Zeitschrift den Zuschlag erhält. Hierbei werden entlang des Einheitsintervalls die Anteile der einzelnen Zeitschriften an der Gesamtzahl der schon verteilten Artikel abgetragen. Der Zufallszahlengenerator erzeugt anschließend eine Zahl zwischen 0 und 1. Somit erfolgt die Vergabe aufgrund der oben definierten Wahrscheinlichkeiten.

4.2.1 Basis-Version

Die Basis-Version des Programms entspricht dem reinen Yule-Simon-Prozess. Es sollte also weitestgehend die in Abschnitt 4.1.1 gefundene analytische Lösung reproduziert werden. Abbildungen 4.1 bis 4.3 zeigen die kumulierten Verteilungsfunktionen für sechs verschiedene Simulationen. Analog zur Darstellungsweise in Kapitel 3 zeigt die x-Achse die Artikelzahl und die y-Achse den Anteil, den Zeitschriften mit mehr als x oder gleich x Artikel an der Gesamtzahl aller Artikel besitzen.

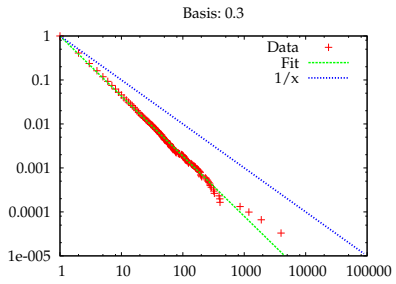


(a) $\alpha = -1.07$; $q = 0.1$; $\alpha_{theo} = -1.11$

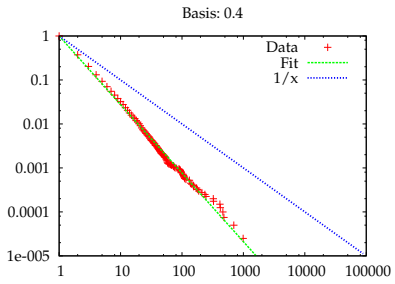


(b) $\alpha = -1.21$; $q = 0.2$;$\alpha_{theo} = -1.25$

Abb. 4.1: Basis Version; $N = 20000$;

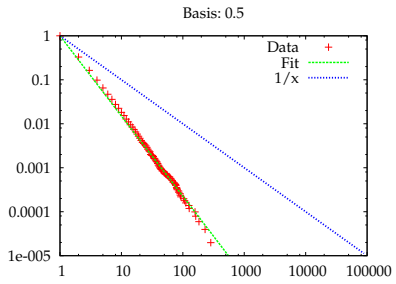


(a) $\alpha = -1.36$; $q = 0.3$; $\alpha_{theo} = -1.36$

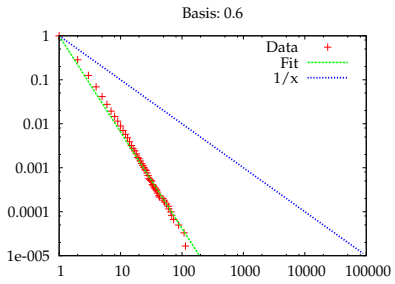


(b) $\alpha = -1.55$; $q = 0.4$;$\alpha_{theo} = -1.66$

Abb. 4.2: Basis Version; $N = 20000$;



(a) $\alpha = -1.80$; $q = 0.5$; $\alpha_{theo} = -2$



(b) $\alpha = -2.18$; $q = 0.6$;$\alpha_{theo} = -2.5$

Abb. 4.3: Basis Version; $N = 20000$

Bei den Simulationen wurden für die Neubesetzungswahrscheinlichkeit q die Werte 0.1, 0.2, 0.3, 0.4, 0.5 und 0.6 gewählt. Zu erkennen ist jeweils ein ausgeprägtes Potenzgesetzverhalten. Leichte Abweichungen hiervon, vor allem bei hohen Artikelzahlen, sind weitestgehend als numerisches Rauschen bzw. numerische Instabilitäten zu bewerten. Die Bildunterschrift gibt neben der Neubesetzungswahrscheinlichkeit q auch den durch Anpassung einer Potenzfunktion berechneten Wert für den Exponenten der Verteilung α sowie den nach Gleichung 4.18 berechneten theoretischen Wert α_{theo} für diesen Exponenten (immer bezogen auf die kumulierte Verteilungsfunktion) an. Zur Orientierung ist in den Abbildungen zusätzlich die Potenzfunktion mit dem Exponenten $\alpha = 1$ eingezeichnet. Die Streuung der Punkte um die ideale Potenzfunktion soll

noch eingehender diskutiert werden: Im Programm wurde ein multiplikativer Zufallszahlengenerator verwendet. Dieser basiert auf der Formel:

$$x_{i+1} = (c * x_i) \bmod p \tag{4.21}$$

Dies bedeutet, dass der jeweils nächste Wert in einer Reihe von Zufallszahlen sich aus dem vorherigen Wert dadurch errechnet, dass der alte Wert mit einer Konstanten c multipliziert wird. Anschließend wird das Resultat modulo einer Konstanten p genommen. Dieses Verfahren benötigt einen vorgegebenen Startwert x_0. Multiplikative Zufallszahlengeneratoren stehen durchaus in der Kritik, Zahlenfolgen von nicht sehr hoher Güte zu liefern. Abhilfe können hier z.T. additive Generatoren schaffen. Diese benötigen jedoch mehr Rechenzeit. In jedem Fall handelt es sich bei den erzeugten Zahlen nicht um echte Zufallszahlen, sondern vielmehr um sogenannte Pseudozufallszahlen, die zwar einer im mathematischen Sinne chaotischen Verteilung folgen, jedoch immer streng deterministisch sind. Das bedeutet unter anderem, dass bei den durchgeführten Simulationen bei gleicher Parameterwahl für den Generator, gleichem q und gleicher Anfangsbedingung für die Zeitschriftenverteilung stets die gleiche Verteilung berechnet wird. Hierbei werden sich mögliche Instabilitäten also in immer gleicher Weise aufschaukeln und verstärken und somit zu den immer gleichen Ausreißern in der simulierten Verteilung führen. Idealerweise sollte das Ergebnis der Verteilung jedoch unabhängig vom Startwert des Zufallszahlengenerators (dem sogenannten Keim) sowie natürlich auch unabhängig von der Anfangsbedingung für die Verteilung der Zeitschriften sein. Um dies im vorliegenden Fall zu erreichen, müssten für jede Verteilung, die simuliert werden soll, eine Vielzahl von Startkonfigurationen gewählt werden (bei Variation des Startwertes des Zufallszahlengenerators und Variation der Anfangsbedingung der Zeitschriftenverteilung) und die vollständige Berechnung durchgeführt werden. Aus der Menge der erhaltenen Verteilungen würde dann eine Wahrscheinlichkeitsverteilung der besten Verteilung bestimmt werden. Dies auch als Monte-Carlo-Verfahren bekannte Vorgehen würde jedoch zu einem enormen Anwachsen des numerischen Aufwands führen, der nur bedingt durch die zu erwartende Verbesserung des Ergebnisses gerechtfertigt werden könnte. Im Rahmen dieser Arbeit wird daher auf die Durchführung dieses Programms verzichtet. Die Streuung der Simulation um die ideale Potenzfunktion wird als tolerierbar aufgefasst. Eine weitere Auffälligkeit ist die mit wachsendem q ebenfalls wachsende Differenz zwischen α und α_{theo}. Dieses Faktum ist weniger auf numerische Unzulänglichkeiten zurückzuführen, als vielmehr auf die Tatsache, dass die Näherung in den Gleichungen 4.9 und 4.11 strikt nur für $q << 1$ gilt. Abgesehen von den geschilderten Problemen können die Ergebnisse der Simulationen der Basis-Version als durchaus zufriedenstellend bezeichnet werden.

4.2.2 Erweiterung 1: Wahrnehmung durch Wissenschaftler

Um die Abweichungen der in Kapitel 3 ausgewerteten Stichproben von einem reinen Potenzgesetz zu erklären, sollen in dieser Arbeit zwei Erweiterungen vom originalen Yule-Simon-Prozess simuliert werden. Die in diesem Abschnitt behandelte Erweiterung 1 behandelt die Wahrnehmung der Bedeutung einer wissenschaftlichen Zeitschrift durch die Wissenschaftler. Während im reinen Yule-Simon-Prozess davon ausgegangen

wird, dass die Bedeutung einer Zeitschrift linear mit der Anzahl der schon enthaltenen Artikel zum Thema wächst, soll hier alternativ der Logarithmus bzw. die Wurzel der Artikelzahl genommen werden. Den Effekt dieser Änderung kann man an Abbildung 4.4 ablesen. Abb. 4.4 beschreibt den Zusammenhang zwischen Artikelzahl und relativer Bedeutung von Zeitschriften in der Wahrnehmung von Wissenschaftlern. Die drei Ansätze sind nicht durch tieferliegende Gründe begründ- oder gar beweisbar. Sie folgen vielmehr der grundlegenden Maxime der Einfachheit der Parametrisierung.

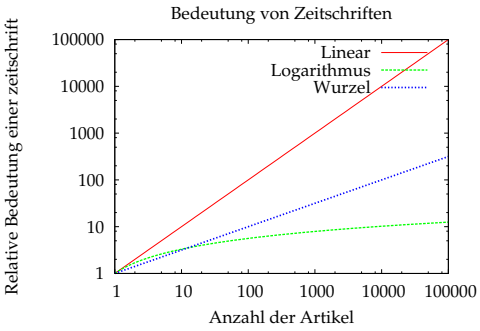


Abb. 4.4: Auswirkungen verschiedener Gewichtungen

Ist der lineare Ansatz schlicht der einfachst mögliche, so lässt sich am ehesten noch für den logarithmischen Ansatz eine Begründung finden. Aus der Physik ist bekannt, dass sich nach dem Weber–Fechner'schen Gesetz die subjektive Stärke eines Sinneseindrucks logarithmisch zur objektiven Intensität des tatsächlichen physikalischen Reizes verhält. Im vorliegenden Fall wäre nun der subjektive Eindruck die Einschätzung eines Wissenschafters zur Bedeutung einer Zeitschrift. Der tatsächliche physikalische Reiz hingegen wäre in dieser Analogie durch die absolute Artikelzahl gegeben. Es soll jedoch betont werden, dass es sich bei den gewählten Parametrisierungen um heuristische Ansätze handelt, für die derzeit kein belastbarer Beweis existiert. Dieser könnte allenfalls durch umfangreiche Befragungen von Wissenschaftlern erbracht werden oder indirekt durch Modellstudien wie den hier vorliegenden. Die Parametrisierung mit der Wurzelabhängigkeit ist in dieser Arbeit lediglich als Mittelweg zwischen linearer und logarithmischer Abhängigkeit gewählt wurden.

Während nun bei einem linearen Anstieg eine Zeitschrift mit 1000 Artikeln auch als 10-mal so bedeutend durch die wissenschaftliche Gemeinschaft empfunden wird wie eine Zeitschrift mit nur 100 Artikeln zum Thema (d.h. es wird auch 10-mal so oft in ihr veröffentlicht), so beträgt der Faktor bei einer Wurzelabhängigkeit nur etwa 3.2. Wird die funktionale Abhängigkeit logarithmisch gewählt, gar nur 1.5 (siehe Abb. 4.4). Zu erwarten ist also, dass bei Variation des funktionalen Zusammenhangs zwischen Artikelzahl und empfundener Bedeutung einer Zeitschrift von „linear“ über die „Wurzel“ hin zum „Logarithmus“ die Ausprägung von Kernzeitschriften, die sehr viele Artikel auf sich vereinigen, immer unbedeutender wird. Es wird also zunehmend mehr Zeitschriften geben, die relativ viele Artikel zum Thema haben. Die Ausprägung der Kernzeitschriften wird entsprechend zurückgehen. In der gewählten Darstellung der kumulativen Verteilungsfunktion entspricht dies einem immer steiler werdendem Abfall der Kurve.

In Formeln beschrieben lauten die jeweiligen Besetzungswahrscheinlichkeiten für die einzelnen Modellversionen nun wie folgt:

In der Basis-Version:

$$p_i = \frac{k_i}{\sum_j k_j} \tag{4.22}$$

In der „Wurzel-Version“:

$$p_i = \frac{\sqrt{k_i}}{\sum_j \sqrt{k_j}} \tag{4.23}$$

In der „Logarithmus-Version“:

$$p_i = \frac{\log k_i + 1}{\sum_j [\log k_j + 1]} \tag{4.24}$$

In der letzten Version muss jeweils zum Logarithmus der Artikelzahl 1 addiert werden. Da $\log 1 = 0$, wäre sonst die Wahrscheinlichkeit zur Besetzung einer Zeitschrift mit einem Artikel gleich Null. In den Abbildungen 4.5 bis 4.16 sind jeweils für den Fall der „Wurzel-Abhängigkeit“ und den Fall der „logarithmischen Abhängigkeit“ für Erstbesetzungswahrscheinlichkeiten von 0.1 bis 0.6 die Simulationen dargestellt. Im Gegensatz zur bisher gewählten alleinigen Darstellung durch ein Log-Log-Diagramm wird hier jedoch zusätzlich die Darstellung in einem Lin-Log-Diagramm gewählt. Bei reiner Betrachtung der Teildiagramme (a), also der Log – Log Diagramme fällt eine Abweichung der Verteilungsfunktion von der reinen Potenzfunktion ins Auge, die schon bei der Diskussion der empirischen Stichproben in Kapitel 3.3 zur Sprache kam. In den dargestellten Simulationen ist diese Abweichung sehr viel extremer zu finden. Die simulierten Verteilungsfunktionen beschreiben keine Gerade im Log-Log-Diagramm, sondern eine mehr oder minder starke konkave Funktion. Diese lässt darauf schließen, dass es sich tatsächlich eher um eine Exponential- denn um eine Potenzfunktion handelt. Daher wurde in der jeweiligen Teilabbildung (b) die Darstellung des Lin-Log-Diagramms gewählt. Hierbei ist die x-Achse linear skaliert, die y-Achse logarithmisch. Ähnlich wie schon bei dem Log-Log-Diagramm gezeigt wurde, dass eine Potenzfunktion in einem solchen Diagramm als Gerade erscheint (Kapitel 2.1), gilt dies für das Lin-Log-Diagramm und eine Exponentialfunktion.

Hierzu sei $f(x)$ eine Exponentialfunktion:

$$f(x) = \exp(\beta x) \tag{4.25}$$

bzw.

$$y = \exp(\beta x) \tag{4.26}$$

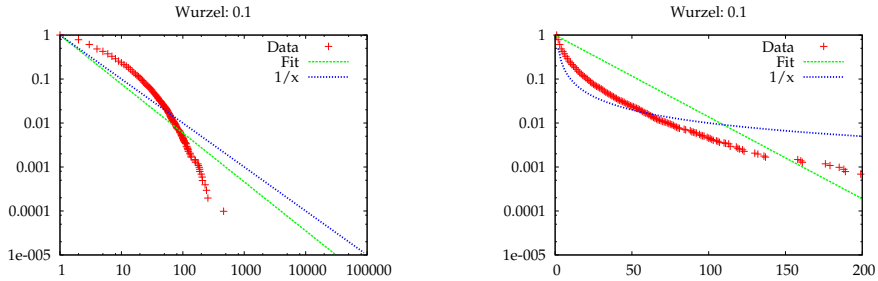
Bei Umskalierung der y-Achse gilt:

$$\tilde{y} = \log(y) \tag{4.27}$$

und somit:

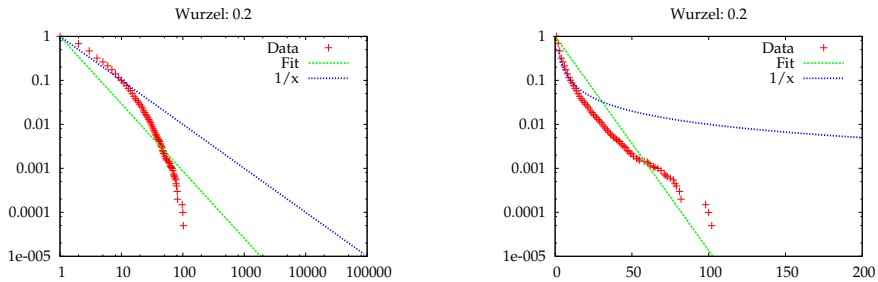
$$\tilde{y} = \log(y) = \log(\exp(\beta x)) = \beta x \tag{4.28}$$

Jede Exponentialfunktion wird in einem Lin-Log-Diagramm somit durch eine Gerade repräsentiert. In den Abbildungen 4.5(b) bis 4.7(b) ist dies nur rudimentär der Fall, in den Abbildungen 4.8(b) bis 4.10(b) schon erheblich deutlicher. Nahezu perfekt ist jedoch die Annäherung in den Abbildungen 4.11(b) bis 4.16(b). Damit kann durch die Abänderung in der Parametrisierung der Wahrnehmung der Bedeutung einzelner Zeitschriften durch die wissenschaftliche Gemeinschaft die tatsächliche funktionale Form des BLS entscheidend beeinflusst werden. Der Streit, ob das BLS nun eine Potenz- oder aber eher eine Exponentialfunktion ist, kann damit als beantwortet angesehen werden. Die Antwort lautet: Mit dem hier vorgestellten Modell können durch eine nachvollziehbare und plausible Abänderung der Parametrisierung des Grundmodells beide Funktionstypen in nahezu beliebiger Weise angenähert werden. In der Praxis wird jedoch in der Regel keine der beiden Funktionen perfekt erreicht. Der geschilderte Sachverhalt kann als eine der Kernaussagen dieser Arbeit gewertet werden. Dieses Ergebnis bestätigt noch einmal, wie wichtig es in diesem Zusammenhang ist, das BLS auf seine Kernaussage zu reduzieren und wie müßig es ist, im Einzellfall bestimmen zu wollen, welcher Funktionstyp das BLS nun am besten beschreibt.



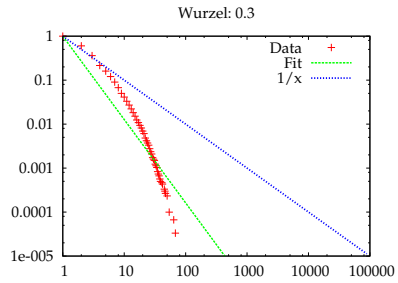
(a) Log - Log - Darstellung; $\alpha = -1.19$; **(b)** Lin - Log - Darstellung $\beta = -0.07$;

Abb. 4.5: Wurzel; $N = 20000$; $q = 0.1$;

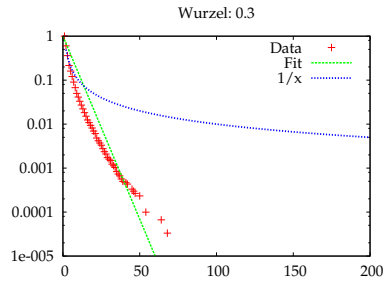


(a) Log - Log - Darstellung; $\alpha = -1.60$; **(b)** Lin - Log - Darstellung $\beta = -0.14$;

Abb. 4.6: Wurzel; $N = 20000$; $q = 0.2$;

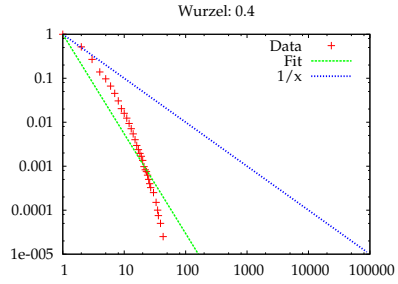


(a) Log - Log - Darstellung; $\alpha = -1.97$;

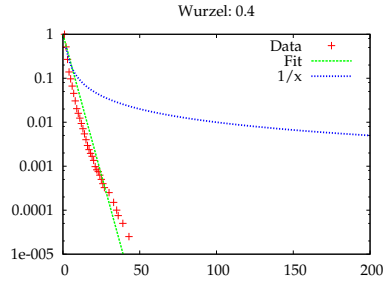


(b) Lin - Log - Darstellung $\beta = -0.22$;

Abb. 4.7: Wurzel; $N = 20000$; $q = 0.3$;

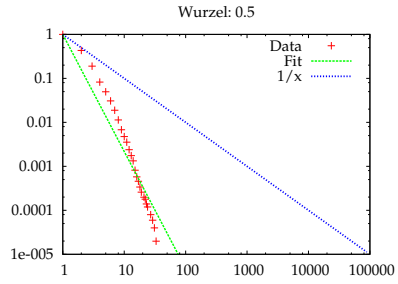


(a) Log - Log - Darstellung; $\alpha = -2.30$;

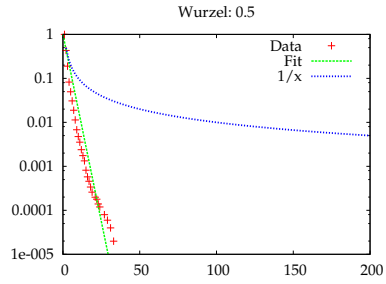


(b) Lin - Log - Darstellung $\beta = -0.35$;

Abb. 4.8: Wurzel; $N = 20000$; $q = 0.4$;

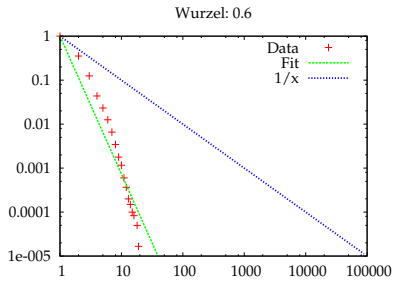


(a) Log - Log - Darstellung; $\alpha = -2.66$;

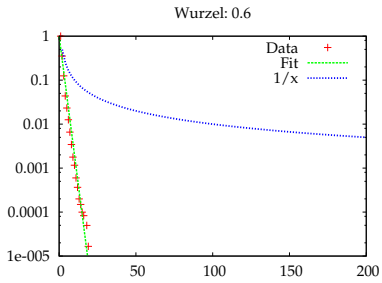


(b) Lin - Log - Darstellung $\beta = -0.46$;

Abb. 4.9: Wurzel; $N = 20000$; $q = 0.5$;

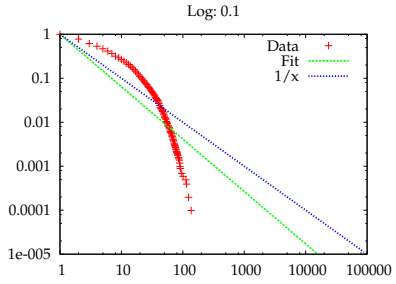


(a) Log - Log - Darstellung; $\alpha = -3.01$;

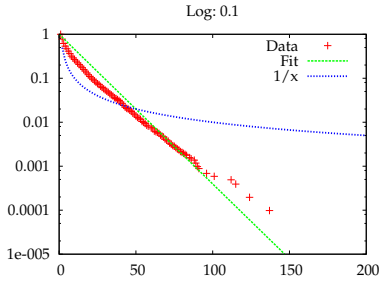


(b) Lin - Log - Darstellung $\beta = -0.61$;

Abb. 4.10: Wurzel; $N = 20000$; $q = 0.6$;

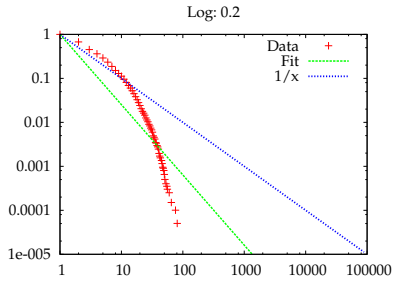


(a) Log - Log - Darstellung; $\alpha = -1.19$;

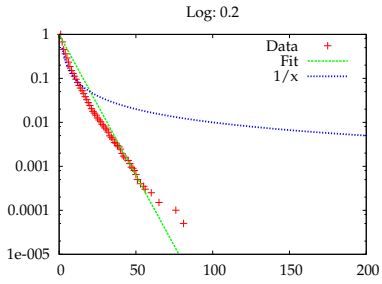


(b) Lin - Log - Darstellung $\beta = -0.07$;

Abb. 4.11: Logarithmus; $N = 20000$; $q = 0.1$;

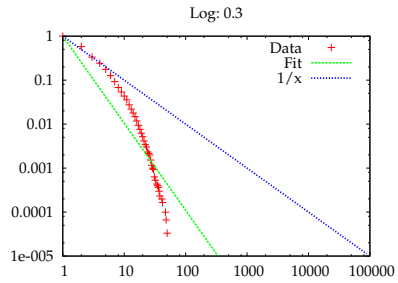


(a) Log - Log - Darstellung; $\alpha = -1.60$;

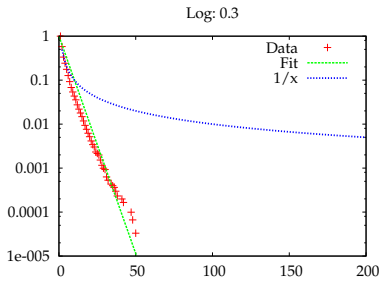


(b) Lin - Log - Darstellung $\beta = -0.14$;

Abb. 4.12: Logarithmus; $N = 20000$; $q = 0.2$;

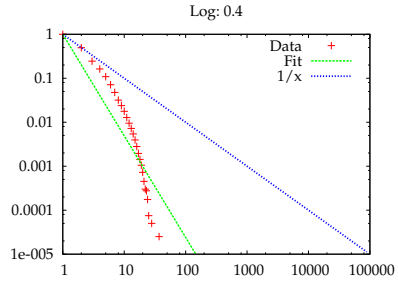


(a) Log - Log - Darstellung; $\alpha = -1.97$;

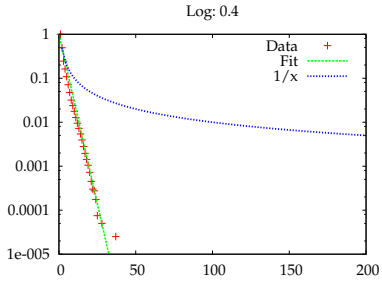


(b) Lin - Log - Darstellung $\beta = -0.22$;

Abb. 4.13: Logarithmus; $N = 20000$; $q = 0.3$;

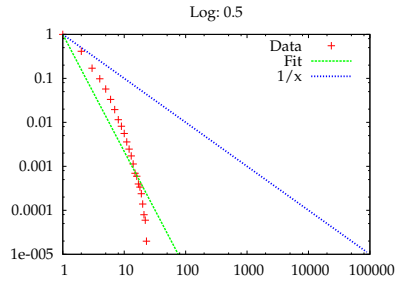


(a) Log - Log - Darstellung; $\alpha = -2.30$;

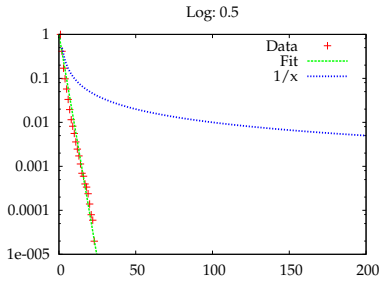


(b) Lin - Log - Darstellung $\beta = -0.35$;

Abb. 4.14: Logarithmus; $N = 20000$; $q = 0.4$;

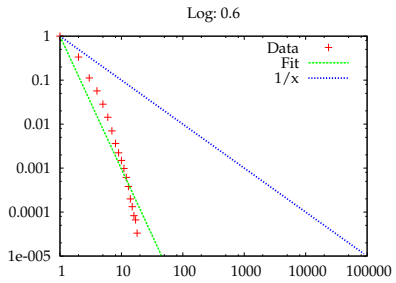


(a) Log - Log - Darstellung; $\alpha = -2.66$;

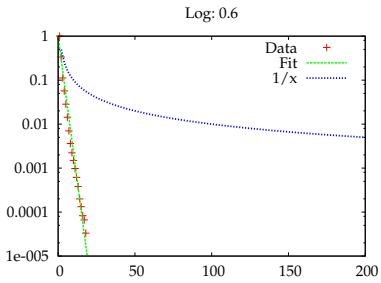


(b) Lin - Log - Darstellung $\beta = -0.46$;

Abb. 4.15: Logarithmus; $N = 20000$; $q = 0.5$;



(a) Log - Log - Darstellung; $\alpha = -3.01$;

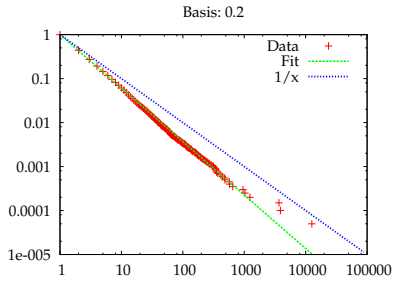


(b) Lin - Log - Darstellung $\beta = -0.61$;

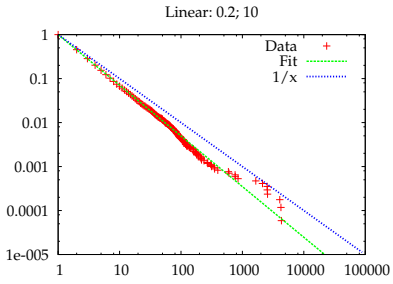
Abb. 4.16: Logarithmus; $N = 20000$; $q = 0.6$;

4.2.3 Erweiterung 2: Ablehnung durch Zeitschriften

Als zweite Erweiterung zum reinen Yule-Simon-Prozess soll dieser dahingehend modifiziert werden, dass seitens der Zeitschriften eine Ablehnung der Artikel möglich wird. Diese wird in der Form berücksichtigt, dass bei steigenden Artikelzahlen, die schon von einer Zeitschrift veröffentlicht wurden, auch die Wahrscheinlichkeit der Ablehnung wächst.

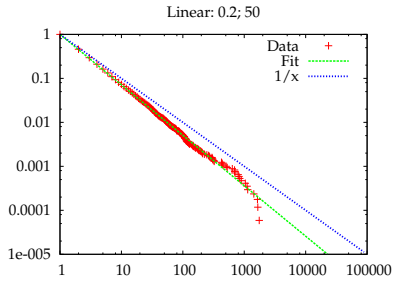


(a) $\alpha = -1.07$; $q = 0.2$;

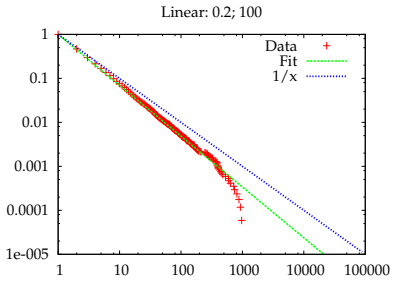


(b) $\alpha = -1.15$; $q = 0.2$; $a = 10$

Abb. 4.17: Basis Version / Linearer Abfall; $N = 20000$;

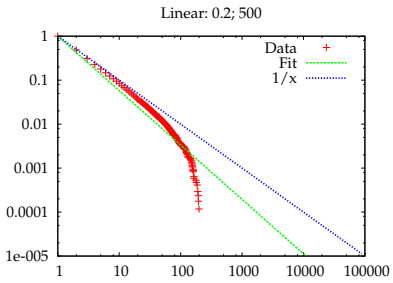


(a) $\alpha = -1.14$; $q = 0.2$; $a = 50$

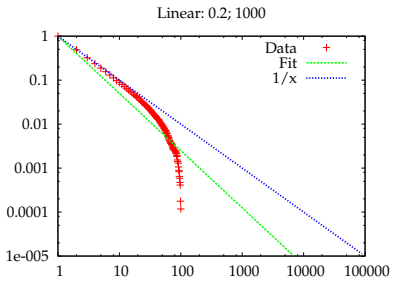


(b) $\alpha = -1.15$; $q = 0.2$; $a = 100$

Abb. 4.18: Linearer Abfall; $N = 20000$;



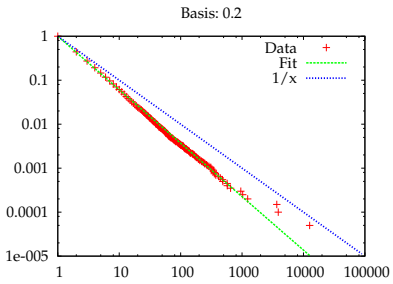
(a) $\alpha = -1.23$; $q = 0.1$; $a = 500$



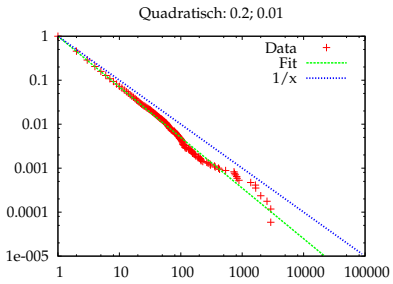
(b) $\alpha = -1.30$; $q = 0.2$; $a = 1000$

Abb. 4.19: Linearer Abfall; $N = 20000$;

Dieses Verhalten soll im Wesentlichen eine Art Sättigung von Zeitschriften beschreiben. Skaliert mit einem freien Parameter, der das relative Aufnahmevermögen der Zeitschriften widerspiegelt, wurde die Ablehnungswahrscheinlichkeit linear, quadratisch und kubisch mit der absoluten Artikelzahl angesetzt. Die Ablehnungswahrscheinlichkeit w_i dafür, dass Zeitschrift i einen Artikel ablehnt, ergibt sich damit zu: $w_i = \frac{ak_i}{N_{ges}}$ beim linearen Fall, $w_i = \frac{a(k_i^2)}{N_{ges}}$ beim quadratischen Fall und $w_i = \frac{a(k_i^3)}{N_{ges}}$ beim kubischen Fall. Hierbei ist a ein freier Parameter, der in der gegebenen Formulierung mit $1/N_{ges}$, also dem Kehrwert der Anzahl aller simulierten Artikel, skaliert wird. k_i ist die Anzahl der Artikel, die von Zeitschrift i schon veröffentlicht wurden. Für jede dieser drei Möglichkeiten wurde jeweils der Fall $q = 0.2$ simuliert.

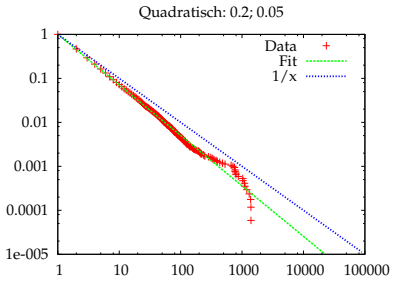


(a) $\alpha = -1.07$; $q = 0.2$;

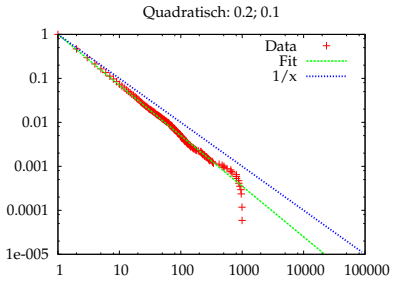


(b) $\alpha = -1.15$; $q = 0.2$; $a = 0.01$

Abb. 4.20: Basis Version / Quadratischer Abfall; $N = 20000$;

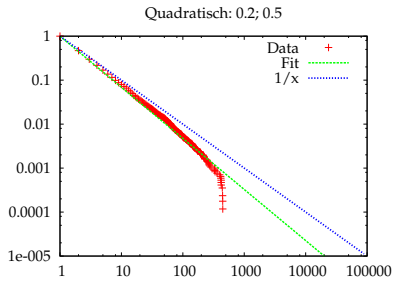


(a) $\alpha = -1.15$; $q = 0.2$; $a = 0.05$

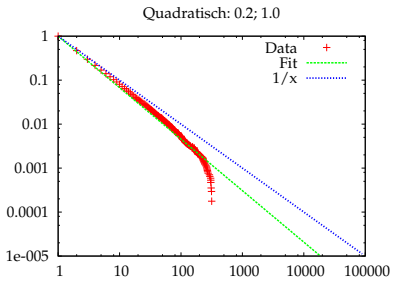


(b) $\alpha = -1.15$; $q = 0.2$; $a = 0.1$

Abb. 4.21: Quadratischer Abfall; $N = 20000$;



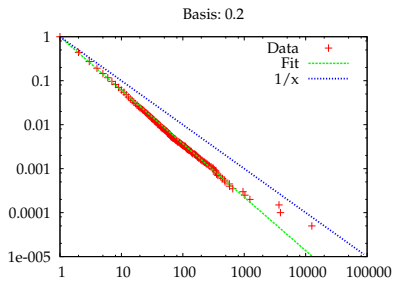
(a) $\alpha = -1.16$; $q = 0.1$; $a = 0.5$



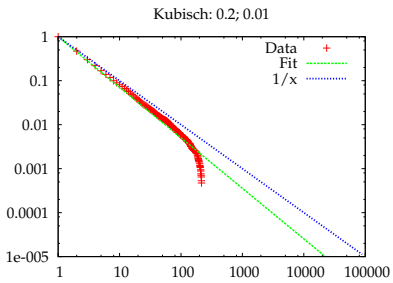
(b) $\alpha = -1.17$; $q = 0.2$; $a = 1.0$

Abb. 4.22: Quadratischer Abfall; $N = 20000$;

Die Resultate (Abb. 4.17 - 4.25) geben ein gutes Bild über die Auswirkungen dieser Variation. Die Möglichkeit der Berücksichtigung von Ablehnungen durch Zeitschriften während des Publikationsprozesses muss im Vergleich zum reinen Yule-Simon-Prozess als wesentliche Erweiterung des Modells betrachtet werden. Dem Ansatz der Parametrisierung folgend sind die Auswirkungen dieses Effekts – in Abhängigkeit des in dieser Modellstudie zunächst willkürlich gewählten freien Parameters zur Ablehnungswahrscheinlichkeit – nur bei den jeweiligen Kernzeitschriften zu betrachten. Es zeigt sich hier ein im Vergleich zum Potenzgesetz verstärkter Abfall der simulierten Verteilungsfunktion. Dieses Charakteristikum konnte schon bei den in Kapitel 3 untersuchten empirischen Stichproben festgestellt werden. Beispiele hierfür sind etwa: laser (WoS/OAIster), music (WoS/OAIster) und investigation (OAIster).

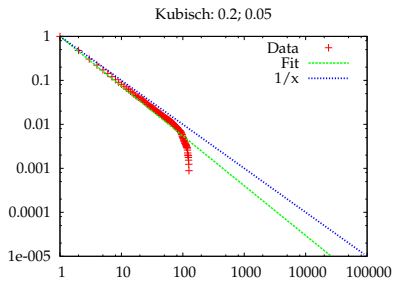


(a) $\alpha = -1.07$; $q = 0.2$;

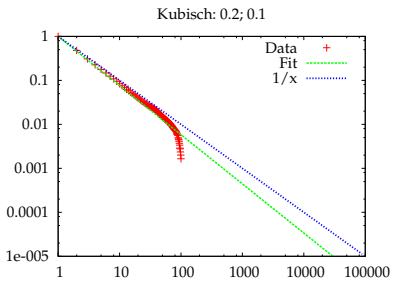


(b) $\alpha = -1.15$; $q = 0.2$; $a = 0.01$

Abb. 4.23: Basis Version / Kubischer Abfall; $N = 20000$;

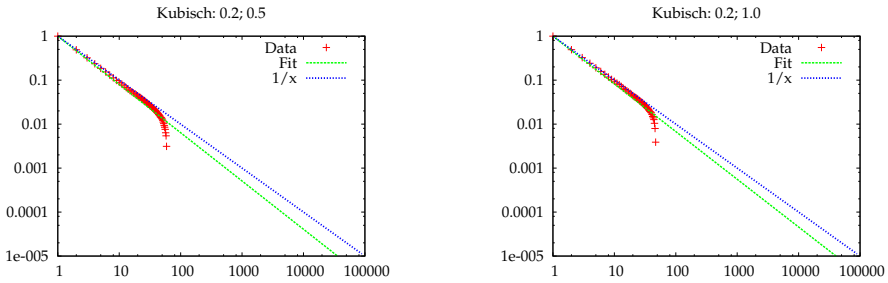


(a) $\alpha = -1.13$; $q = 0.2$; $a = 0.05$



(b) $\alpha = -1.11$; $q = 0.2$; $a = 0.1$

Abb. 4.24: Kubischer Abfall; $N = 20000$;



(a) $\alpha = -1.10; q = 0.1; a = 0.5$ **(b)** $\alpha = -1.08; q = 0.2; a = 1.0$

Abb. 4.25: Kubischer Abfall; $N = 20000$;

4.3 Vergleich der empirischen und simulierten Daten

Das vorgestellte Modell, mit dem die in den letzten Abschnitten vorgestellten Simulationen durchgeführt wurden, beschreibt den Veröffentlichungsprozess durch Wissenschaftler auf sehr rudimentäre und vereinfachte Art und Weise. Ein Vergleich mit den in Kapitel 3 präsentierten tatsächlich gefundenen Verteilungsfunktionen kann und soll daher nur eher qualitativer Natur sein. Wollte man das Verhalten von Wissenschaftlern eines klar definierten Forschungsbereichs detaillierter modellieren, so wären sicherlich weit umfangreichere Daten über die Publikationslandschaft in diesem speziellen Bereich vonnöten, die ihrerseits Eingang finden müssten in ein sehr viel umfangreicheres Modell.

4.3.1 Qualitativer Vergleich

Der qualitative Vergleich von empirischen Stichproben und Simulationen soll sich in erster Näherung auf den Exponenten der Verteilungsfunktion und in zweiter Näherung vornehmlich auf den in vielen Fällen charakteristischen starken Abfall bei hohen Artikelzahlen beschränken. So zeigen – ohne dass hier die Bilder zum wiederholten Male gezeigt werden sollen – die Ergebnisse der Simulationen in den letzten Abschnitten, dass das Modell grundsätzlich in der Lage ist, die Stichproben aus Kapitel 3 gut zu reproduzieren. Über die Freiheiten, die das Basismodell bzw. die vorgestellten beiden Erweiterungen zur Steuerung der Form der Verteilungsfunktion bieten, lassen sich die Simulationen im Prinzip beliebig gut an die Stichproben anpassen.

4.3.2 Detaillierter Vergleich anhand ausgewählter Beispiele

In diesem Abschnitt soll anhand von drei Stichproben aus dem Web of Science detaillierter überprüft werden, inwieweit das numerische Modell in der Lage ist, diese zu reproduzieren. Hierzu wird aus Kapitel 3 pro Modellversion jeweils eine Stichprobe ausgewählt, die mit Hilfe des in diesem Kapitel definierten und getesteten Modells simuliert werden. Alle drei Beispiele werden aus den Stichproben des WoS ausgewählt, da vor allem die schon diskutierte Möglichkeit der Dubletten, wie sie in OAIster vorkommen können, nicht von dem Modell repräsentiert werden können. Im Einzelnen

handelt es sich um die Stichproben zu den Suchworten „revolution“ (für das Basismodell), „phase transition“ (für die Erweiterung 1) und „laser“ (für die Erweiterung 2). Die Anzahl der simulierten Artikel wurde jeweils dem Umfang der Stichproben angepasst. Im Fall „revolution“ wurden 10.000, im Fall „phase transition“ 25.000 und im Fall „laser“ ebenfalls 25.000 Artikel simuliert. In allen drei Fällen, die in den Abbildungen 4.26 bis 4.28 dargestellt werden, konnte eine sehr gute Übereinstimmung zwischen empirischer Stichprobe und Simulation erreicht werden.

Modellversion: Basis; Stichprobe: „revolution“ aus WoS

Im ersten Fall (Abb. 4.26) wurde das reine Basismodell verwendet. Mit dem Wert $q = 0.3$ konnte hier eine gute Übereinstimmung zwischen Modell und Stichprobe erreicht werden. Einziges Manko ist die starke Abweichung der beiden Topzeitschriften. Während die höchstbesetzte Zeitschrift in der Stichprobe etwa 200 Artikel enthält, sind dies im Modell immerhin etwa 1000. Diese enorme Diskrepanz muss auf die weiter oben diskutierten möglichen Instabilitäten zurückgeführt werden. Man muss sich bei einer doppelt logarithmischen Darstellung immer darüber im Klaren sein, dass nicht unbedeutende Abweichungen leicht unterschätzt werden. Bis auf die beschriebene Diskrepanz ist die Übereinstimmung jedoch überzeugend. Zur Beurteilung der Diskrepanz zwischen Stichprobe und Simulation sei hier auch noch zusätzlich auf Kapitel 5 verwiesen.

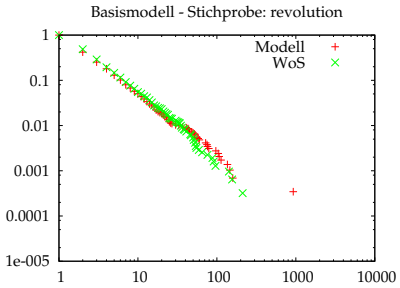


Abb. 4.26: Vergleich: Basismodell - revolution (WoS)

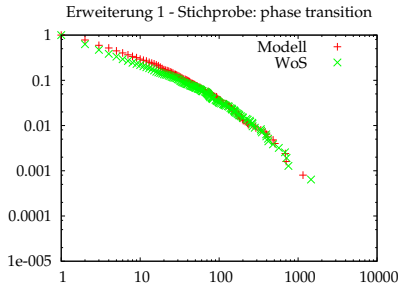


Abb. 4.27: Vergleich: Erweiterung 1 - phase transition (WoS)

Modellversion: Erweiterung 1; Stichprobe: „phase transition“ aus WoS

Bei der Stichprobe „phase transition“ wurde aufgrund der Form der Verteilungsfunktion angenommen, dass die Erweiterung 1, d.h. die Modifizierung der Wahrnehmung der Bedeutung von Zeitschriften durch Wissenschaftler, zu einem guten Ergebnis führen könnte. Zunächst wurde jedoch weder mit der „Wurzel-Version“ noch mit der „Logarithmus-Version“ ein zufriedenstellendes Ergebnis erreicht. Daher wurde der Wurzelansatz in der Form erweitert, dass neben 0.5 als Exponent beliebige Exponenten kleiner 1 zugelassen wurden. Auf diese Weise konnte mit dem Wert 0.7 als Exponent und $q = 0.05$ eine gute Übereinstimmung erreicht werden. Der Übergang vom Exponenten 0.5 (also der Wurzel) hin zu beliebigen Exponenten kleiner 1 wird dabei nicht als neuer Ansatz verstanden, sondern nur als Modifizierung der „Wurzel-Version“ des

Modells. Zu dieser Modellversion ist zu bemerken, dass aufgrund der Analogie zum Weber-Fechner'schen Gesetz die „Logarithmus-Version" aus physikalischer Sicht die befriedigenste ist. Dass diese hier nicht direkt zum Erfolg führt, ist darauf zurückzuführen, dass auch hier wieder über ein statistisches Ablehnverhalten seitens der Zeitschriften gemittelt werden müsste.

Modellversion: Erweiterung 2; Stichprobe: „laser" aus WoS

Wie schon bei Fall 2 wurde auch hier per Augenschein die Entscheidung getroffen, Erweiterung 2 als Referenz-Modell zur Reproduktion der Verteilungfunktion heranzuziehen. Mit der linearen Sättigung, $a = 50$ und $q = 0.1$ war auch hier das Ergebnis zufriedenstellend.

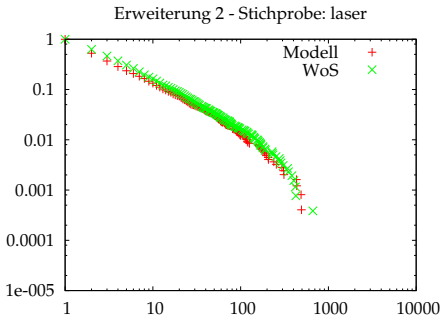


Abb. 4.28: Vergleich: Erweiterung 2 - laser (WoS)

Die Anpassungen wurden jeweils per Hand und Augenschein, d.h. durch Testen einiger weniger Parameterkonstellationen vorgenommen. Es handelt sich also nicht um eine Optimierung mit echter Fehlerberechnung etwa im Sinne der Berechnung der kleinsten Fehlerquadrate. Ein solches Verfahren könnte zwar zu besseren Ergebnissen führen, ginge aber bei Berücksichtigung des gesamten Parameterraumes auch mit einem enormen numerischen Aufwand einher. Zusammenfassend läßt sich feststellen, dass das Modell durchaus in der Lage ist, auch Details und gerade auch systematische Abweichungen von einer idealen Potenzfunktion zu reproduzieren, wenn die entsprechenden Erweiterungen des Yule-Simon-Prozesses herangezogen werden.

Kapitel 5

Alternative Darstellungen

Nach der bisherigen Präsentation der Daten ist selbst dem mathematisch Geübten und mit für im Bibliothekswesen eher ungewöhnlichen Konstrukten wie Verteilungsfunktionen vertrauten Leser klar geworden, dass eine umfassende und breite Analyse des Publikationswesens auf dem bisher beschrittenen Wege nicht möglich ist. Auf der einen Seite ist es sicherlich nötig – ja unerlässlich –, Verteilungen im Detail zu betrachten, um grundlegende Strukturen und Mechanismen bei der Verteilung von Artikeln auf Zeitschriften zu erkennen. Auf der anderen Seite ist es schlicht unmöglich, dies für eine größere Zahl von Stichproben durchzuführen und dennoch den Überblick zu behalten. Im Folgenden soll es daher darum gehen, anhand der bisher präsentierten Ergebnisse andere, alternative Darstellungen von Konzentration und Diversität vorzustellen und kurz zu diskutieren.

Entscheidend ist dabei, die gesamte Information über eine Verteilung in einer aussagekräftigen Zahl gleichsam zu kondensieren. Dies ist im Rahmen dieser Arbeit schon zum Teil geschehen. Die Angabe der Steigung einer angepassten Potenz- bzw. Exponentialfunktion erfolgte genau in dieser Absicht. Bei Betrachtung der Abbildungen ist jedoch auch klar geworden, dass diese Steigung zum Teil ein sehr schlechter und nur angenäherter Weg, die jeweilige Verteilung zu charakterisieren, sein kann. Es sollen in diesem Kapitel nun die Lorenz-Kurve, der damit eng verwandte Gini-Koeffizient, die Shannon-Entropie sowie das Theil-Maß vorgestellt und angewendet werden. Im Bereich der Bibliometrie wurden die hier benutzten Maße schon von [*Heinz*, 2006] zur Bestimmung der Internationalität der Forschung benutzt.

5.1 Alternativen zur Messung von Diversität und Konzentration

Diversität bzw. Konzentration soll in unserem Fall natürlich auf die „Population“ der Zeitschriften und deren Artikel angewendet werden. Die Wortwahl unterstreicht die Verwandtschaft des Themenkomplexes mit der Biologie. Wird dort die Artenvielfalt von Flora und Fauna untersucht, so soll hier die Interdisziplinarität bzw. die Konzentration der Verteilung auf etwaige Kernzeitschriften quantifiziert werden.

Lorenz-Kurve und Gini-Koeffizient

Zunächst soll nun die sogenannte Lorenz-Kurve eingeführt werden. Diese ist besonders dazu geeignet, ein Ungleichgewicht bzw. eine Konzentration in einem gegebenen Datensatz graphisch darzustellen. Hierzu sei eine Stichprobe der Mächtigkeit n gegeben durch einen Vektor $\mathcal{X} = (x_1, x_2, \ldots, x_n)$, wobei gilt: $x_1 \leq x_2 \leq \ldots. \leq x_n$. Nun ist der Vektor der normierten Partialsummen gegeben durch $\mathcal{Y} = (y_1, y_2, \ldots, y_n)$. Hierbei gilt:

$$y_i = \frac{\sum_{j=1}^{i} x_j}{\sum_{j=1}^{n} x_j} \tag{5.1}$$

Die Zahlenpaare $(y_i, \frac{\sum_{j=1}^{i} j}{\sum_{j=1}^{n} j})$ ergeben dann die Lorenz-Kurve.

Der Gini-Koeffizient ist definiert als das Doppelte der Fläche, die von der Lorenz-Kurve und der Hauptdiagonalen im Koordinatenkreuz eingeschlossen wird. Anschaulich gilt: Je größer diese Fläche ist, desto stärker ist die Verteilung auf einige wenige Cluster konzentriert. Ist die Lorenz-Kurve gleich der Hauptdiagonalen, liegt absolute Gleichverteilung vor. Ist sie durch eine Stufenfunktion gegeben mit den Funtionswerten 0 für $x < 1$ und 1 für $x = 1$ so ist die Verteilung auf einen einzigen Punkt konzentriert. Bezeichnet $\mathcal{L}(x)$ also die Lorenz-Kurve, so gilt für den Gini-Koeffizienten:

$$\mathcal{G} = 1\text{–}2 \int_0^1 \mathcal{L}(x)dx \tag{5.2}$$

Der Gini-Koeffizient ist ein beliebtes Maß in der Volkswirtschaftslehre, um beispielsweise das Ungleichgewicht von Einkommensverteilungen zu bestimmen [*Wolf*, 1997].

Shannon-Entropie

Die Shannon-Entropie ist ein Maß aus der Informationstheorie zur Bestimmung des Informationsgehalts eines Zeichens in einer Folge von Zeichen. Information kann dabei nach Shannon als beseitigte Unbestimmtheit bezeichnet werden. In unserem Zusammenhang gilt dann: Je interdisziplinärer die Verteilung von Artikeln zu einem bestimmten Suchwort ist, desto weniger genau kann bei einem gegebenen Artikel vorhergesagt werden, aus welcher Zeitschrift er stammt. Während der Gini-Koeffizient mit wachsender Konzentration der Verteilung größer wird, nimmt die Entropie im gleichen Fall ab. Bei wachsender Konzentration kann also ein beliebiger Artikel mit immer höherer Wahrscheinlichkeit einer Kernzeitschrift zugeordnet werden.

Die Entropie einer Verteilung (als Zeichenkette aufgefasst) ist definiert als [*Shannon*, 1976]:

$$\mathcal{H}(x) = -\sum_{i=1}^{n} x_i \log_2(x_i) \tag{5.3}$$

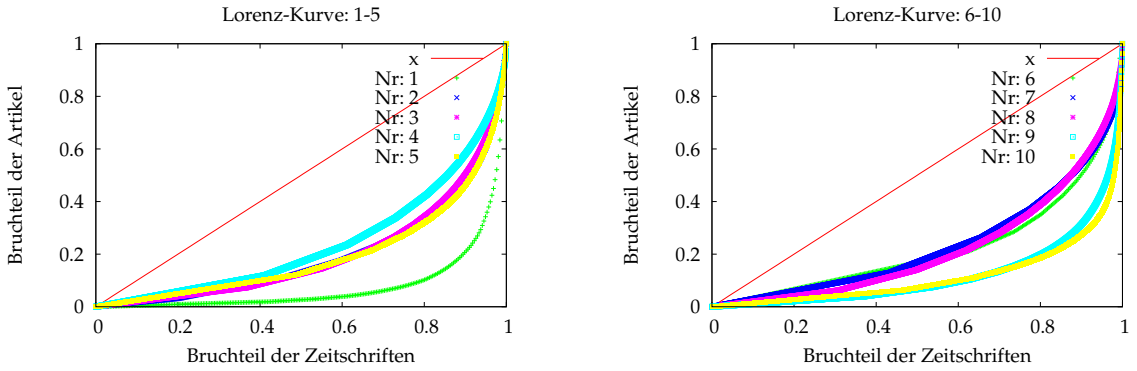
Theil-Maß

Das Theil-Maß basiert ebenfalls auf dem Entropieansatz. Es setzt die oben definierte Entropie $\mathcal{H}(x)$ in Beziehung zur maximal möglichen Entropie, die im Falle einer absoluten Gleichverteilung der Verteilung vorläge. Es ist definiert durch [*Theil*, 1967]:

$$\mathcal{T}(x) = \log_2(n) - \mathcal{H}(x) = \log_2(n) + \sum_{i=1}^{n} x_i \log_2(x_i) \tag{5.4}$$

Diese drei Maße ermöglichen – ähnlich wie die Charakterisierung der Verteilung über eine angepasste Potenz- bzw. Exponentialfunktion – eine kompakte Beschreibung einer im Detail möglicherweise sehr detailreichen Verteilung. Darüber hinaus handelt es sich um - vor allem, aber nicht nur in der Ökonomie – anerkannte und erprobte Ansätze und bietet damit die Möglichkeit, bibliometrische Ergebnisse in einer für nicht bibliothekarisch erfahrene Fachwissenschaftler nachvollziehbaren Weise zu präsentieren. Die drei vorgestellten Maße sollen nun zunächst für die ersten 10 Stichproben zum WoS aus Kapitel 3 berechnet werden. Dies soll einen Überblick über die Spanne der möglichen Werte und deren Beziehung untereinander ermöglichen, bevor die drei Fälle des detaillierten Vergleichs zwischen empirischen Stichproben und Simulationen aus Kapitel 4.3.2 ebenfalls anhand der drei Maße überprüft werden.

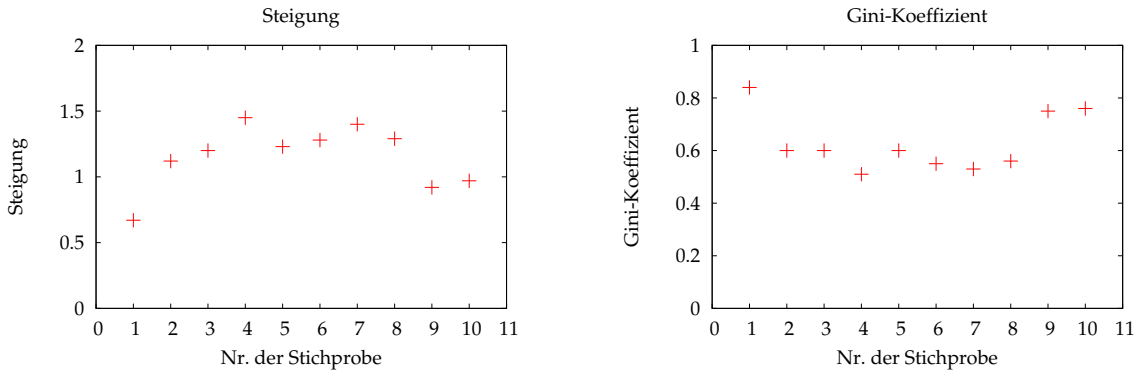
Zunächst werden in Abb. 5.1 die Lorenz-Kurven der 10 fraglichen Fälle dargestellt. Die Zuordnung von Stichprobennummer und Suchbegriff sind der Tabelle 3.1 zu entnehmen.



(a) Datensätze 1-5 **(b)** Datensätze 6-10

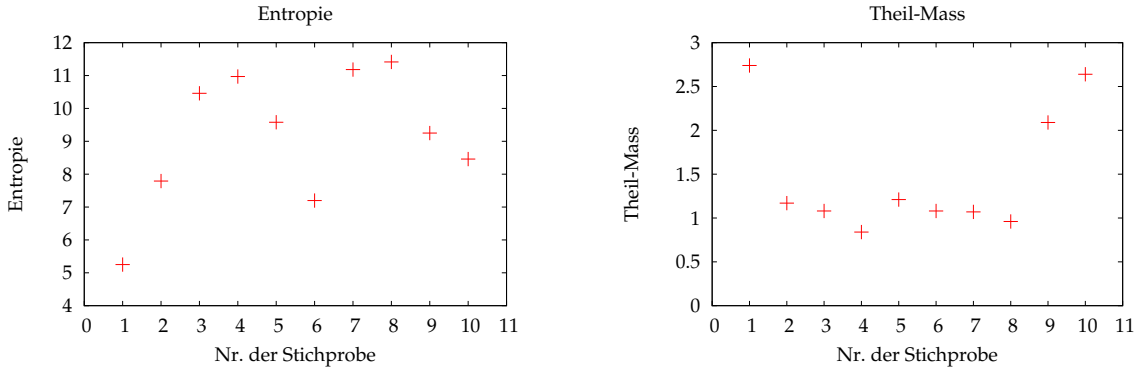
Abb. 5.1: Lorenz-Kurven WoS; Stichproben 1-10

Entsprechend kann nun anhand der Lorenz-Kurve der Gini-Koeffizient berechnet werden. Zusätzlich werden in den Abb. 5.2 und 5.3 die Steigung der angepassten Potenzfunktion, die berechnete Entropie sowie das Theil-Maß angegeben.



(a) Steigung **(b)** Gini-Koeffizient

Abb. 5.2: Steigung und Gini-Koeffizient der 10 Stichproben



(a) Entropie **(b)** Theil-Maß

Abb. 5.3: Entropie und Theil-Maß der 10 Stichproben

Abb. 5.2 und 5.3 zeigen sehr deutlich, dass die vier Maße zur Beschreibung von Konzentration bzw. Diversität der Verteilungen in etwa die gleiche Information wiedergeben. Hierbei können die Maße in zwei Gruppen unterteilt werden: Der Gini-Koeffizient und das Theil-Maß nehmen bei wachsender Konzentration zu und bei wachsender Diversifizierung ab. Bei der Steigung und der Entropie ist das Verhalten definitionsgemäß invers. Die Sensitivität der verschiedenen Maße auf Unterschiede in den Verteilungen ist dabei durchaus differenziert zu betrachten. Man beachte etwa, dass sich die Stichproben 1, 9 und 10 beim Theil-Maß sehr viel stärker von allen anderen Stichproben unterscheiden als die anderen drei Maße dies erwarten ließen.

Nun sollen die drei hier vorgestellten Maße noch dazu verwendet werden, den detaillierten Vergleich aus Kapitel 4.3.2 zu erweitern. Tabelle 5.1 stellt dazu den Gini-Koeffizienten, die Entropien und die Theil-Maße der drei Fälle aus 4.3.2 für Stichprobe und Simulation gegenüber. Obgleich hier natürlich Vergleichsdaten für die Abschätzung der Güte der Übereinstimmung zwischen Stichprobe und Simulation fehlen, ist durchaus eine erfreulich gute Korrespondenz der Werte festzustellen.

Nach der vorhandenen Datenlage lässt sich abschließend nicht beurteilen, welches Maß nun am besten geeignet ist, die Verteilungen auf kompakte Art und Weise zu beschreiben. So wird z.B. ein einzelner Ausreißer im Bereich der Kernzeitschriften auf

den Gini-Koeffizienten immer eine stärkere Wirkung haben als auf die Steigung (so wie sie hier berechnet wurde). Es wird vielmehr also die geeignete Wahl des Maßes mit der speziellen Anforderung an die Beurteilung der Verteilung verknüpft sein.

Suchbegriff	Gini		Entropie		Theil - Maß	
	Stichp.	Sim.	Stichp.	Sim.	Stichp.	Sim.
revolution	0.57	0.60	10.41	9.86	1.18	1.71
phase transition	0.82	0.78	8.02	8.24	2.51	2.1
laser	0.75	0.76	9.25	9.06	2.09	2.2

Tab. 5.1: Tabellarische Darstellung der untersuchten Stichproben und Simulationen

Kapitel 6

Diskussion und Ausblick

Im Rahmen der vorliegenden Arbeit konnte anhand von Stichproben aus der bibliographischen Datenbank Web of Science und der Metasuchmaschine für Open Access Dokumente OAIster gezeigt werden, dass das Bradford Law of Scattering für die Verteilung von wissenschaftlichen Publikationen bzw. Artikeln auf Zeitschriften bzw. online Repositorien eine gute erste Näherung ist. Erste Näherung deshalb, weil das von Bradford postulierte Potenzgesetzverhalten die qualitative Verteilung sicherlich recht gut beschreibt, jedoch die mannigfaltigen, keinesfalls unerheblichen Abweichungen, die die gewonnen Stichproben von einem idealen Potenzgesetz zeigen, nicht zu erklären vermag. Um die charakteristischen Abweichungen der empirischen Stichproben vom Potenzgesetz zu untersuchen, wurden ausgehend vom klassischen Yule-Simon-Prozess, numerische Simulationen durchgeführt. Hierbei wurde zunächst der Yule-Simon-Prozess, der an sich noch analytisch lösbar ist und als Modell für das Publikationsverhalten von Wissenschaftlern gelten kann, als markov'scher Zufallsprozess kodiert. Die analytisch gefundene Lösung des Prozesses konnte damit leicht bestätigt werden. Als Erweiterung des klassischen Yule-Simon-Prozesses wurde das Modell in zwei wesentliche Richtungen modifiziert: Während sich der Yule-Simon-Prozess lediglich am klassischen preferential attachment orientiert, d.h. Wissenschaftler wählen Zeitschriften proportional zu der schon vorhandenen Zahl von Artikeln zu einem Thema in dieser Zeitschrift aus (wobei die Zeitschrift jeden Artikel akzeptiert), wurde hier die Möglichkeit der Ablehnung eines Artikels durch eine Zeitschrift modelliert. Zudem wurden zwei weitere funktionale Abhängigkeiten bei der Wahrnehmung von Zeitschriften durch Wissenschaftler in Betracht gezogen. Beim preferential attachment besteht ein linearer Zusammenhang zwischen der Anzahl der Zeitschriften zu einem Thema und der Wahrscheinlichkeit, dass ein Wissenschaftler in dieser Zeitschrift veröffentlicht. Als Alternativen wurden hier in Betracht gezogen: die Wahrscheinlichkeit einer Neupublikation als Logarithmus bzw. als Wurzel der Anzahl der schon vorhandenen Artikel. Mit Hilfe dieser beiden heuristischen Annahmen konnten charakteristische Abweichungen vom Potenzgesetz, wie sie die untersuchten Stichproben aufweisen, erklärt werden. Während sich ein relativ starker Abfall der Stichprobe bei hohen Artikelzahlen im Vergleich zum Potenzgesetz vornehmlich durch eine Ablehnung von Artikeln durch die Zeitschriften erklären lässt, muss – im Rahmen dieser Arbeit – ein generell schnellerer Abfall der Stichprobe über ihren gesamten Verlauf eher auf eine nicht lineare funktionelle Abhängigkeit der Publikationswahrscheinlich-

keit von der vorgefundenen Artikelzahl zurückgeführt werden. Die beiden Ansätze für die Erweiterungen sind derzeit als heuristisch und intuitiv zu bewerten und können nicht in strikter Weise aus einer zugrundeliegenden Theorie entwickelt bzw. durch belastbares Zahlenmaterial untermauert werden. Sie sind jedoch hinreichend plausibel und nahe liegend, um als ernstzunehmende Kandidaten zur besseren Beschreibung des Publikationsprozesses gelten zu können. So ist es z.B. unmittelbar einsichtig, dass die Topzeitschriften eines Themas es sich eher leisten können, Artikel abzulehnen bzw. dies aufgrund der hohen Zahl der eingereichten Artikel sogar müssen. In welcher Form diese Ablehnwahrscheinlichkeit nun genau parametrisiert werden kann, ist derzeit nicht sicher. Die in dieser Arbeit vorgestellten Stichproben sowie die durchgeführten Simulationen legen jedoch nahe, dass der gewählte Ansatz als guter Einstieg dienen kann. Man könnte sich jedoch auch hier weitere Möglichkeiten, wie etwa die zweite und dritte Vorlage abgelehnter Artikel, vorstellen. Diese Einflüsse können jedoch als Effekte höherer Ordnung angesehen werden und sollen daher zunächst außer Acht gelassen werden.
Unklarer ist die Situation bei der Parametrisierung der Wahrnehmung der Bedeutung von Zeitschriften durch die wissenschaftliche Gemeinschaft. Verläuft diese linear mit der Artikelzahl oder etwa logarithmisch? Die Antwort muss wahrscheinlich lauten: Beides und keines von beiden! Realistischerweise müsste auch bei der Beschreibung dieses Zusammenhangs ein stochastischer Anteil berücksichtigt werden. Die wissenschaftliche Gemeinschaft – auch die Teilgruppe, die auf einem klar definierten Teilgebiet forscht – kann niemals als homogen angesehen werden. Dies würde jedoch ein einheitlicher funktionaler Zusammenhang voraussetzen. Persönliche Vorlieben für Zeitschriften mischen sich mit Vorlieben und anderweitig motivierten Präferenzen von Teilgruppen. Editorenschaft und Gutachtertätigkeit mögen Aspekte sein, die bei der Auswahl einer Zeitschrift für die Publikation eines Artikels eine Rolle spielen. Ob sich diese unterschiedlichen Einflüsse am Ende mit einem einfachen funktionalen Zusammenhang abbilden lassen, kann im Rahmen dieser Arbeit nicht beantwortet werden. Die Grundidee kann jedoch als richtig angenommen werden. Zur Klärung der Details wären weitere Untersuchungen notwendig.
Mit der Modellierung derartiger Prozesse eröffnet sich also ein ungemein spannendes und zukunftsweisendes Themengebiet, dem die Bibliotheks- und Informationswissenschaft derzeit möglicherweise noch nicht die notwendige Beachtung schenkt. Die Verfügbarkeit und Handhabbarkeit immer größer werdender Datenmengen, sei es für den kommerziellen Zeitschriftenmarkt oder aber im OA-Bereich, wird in Zukunft stetig zunehmen. Wie in dieser Arbeit demonstriert wurde, ist es relativ leicht entsprechend große Datenmengen, die die Anwendung statistischer Methoden erlauben, zusammenzutragen. Dies eröffnet die Chance in nie gekannter Weise empirische Stichproben aus großen Datenbanken mit physikalisch und statistisch begründeten Computersimulationen zu vergleichen. Die Methoden, die die statistische Physik bzw. die Soziophysik für derlei Untersuchungen bereitstellt, haben in den letzten Jahren und Jahrzehnten, nicht zuletzt durch die immer leichter verfügbare hohe Rechenleistung moderner Computer und die damit verbundene Umsetzbarkeit von analytisch nicht mehr zu lösenden Modellen, enormen Zuwachs erhalten. Während sich die vorliegende Arbeit eines relativ einfachen Modells auf Basis der Mastergleichung bedient, sind durchaus komplexere Ansätze im Rahmen der Physik der komplexen Systeme

sorgfältig formuliert und daher mit begrenztem Aufwand auf Fragestellungen, wie sie in der Bibliometrie zu finden sind, übertragbar. Diese Ansätze finden sich z.B. in den noch relativ jungen Disziplinen wie der Verkehrsphysik, der Ökonophysik und der Soziophysik. Der mathematische Apparat, der dort zur Anwendung kommt, erstreckt sich von den einfacheren Konzepten des random walks und der Brown'schen Bewegung über Markov-Prozesse im Allgemeinen bis hin zu anspruchsvolleren Ansätzen wie etwa der Fokker-Planck-Gleichung, der Theorie stochastischer Differentialgleichungen der Lotka-Volterra-Gleichung und der Programmierung von agentenbasierten Modellen oder zellulären Automaten. Wie hier im Kapitel 2 am Beispiel der Analogie zwischen dem BLS und der Clusterbildung von Städten vorgeführt wurde, ist die Hauptaufgabe bei der Anwendung solcher Ansätze in der Bibliometrie der Transfer-Prozess. Es muss also das Rad nicht jeweils neu erfunden werden, sondern es müssen erfolgreiche Konzepte mittels Analogieüberlegungen übersetzt werden. Dieses Vorgehen steht ganz in der Tradition von [*de Solla Price*, 1963], der vorschlug, die von ihm postulierte Big Science mit Hilfe der Thermodynamik berechenbar zu machen.
Doch wo liegen die möglichen praktischen Anwendungen einer so betriebenen Bibliometrie?
Für die bibliothekarische Praxis sei hier nur eine - recht konventionelle - Anwendung genannt: Der Zeitschriftenpublikationsmarkt, wie er sich heute darstellt, hat, nicht zuletzt durch den hinzugekommenen Sektor der elektronischen Zeitschriften, eine Vielfalt und Komplexität sowie eine Größe erreicht, die noch vor 10 – 20 Jahren unvorstellbar war. Hinzu hat sich in den letzten Jahren der ebenfalls enormen Wachstumsraten unterliegende Bereich der Open Access Publikationen gesellt. Geht man von den seit langer Zeit geltenden Wachstumsraten für wissenschaftliche Kommunikation und Publikation aus, so wird sich entsprechend einem exponentiellen Wachstum der betrachtete Bereich weiter mit einer Rate von etwa 10 - 15 Jahren verdoppeln [*de Solla Price*, 1963]. Dieser Prozess geht für Bibliotheken mit enormen Chancen, aber auch Risiken einher. In jedem Fall stellt er jedoch eine Herausforderung dar, die mit neuen Konzepten und Methoden der Bibliotheks- und Informationswissenschaft beantwortet werden muss.

Nicht wenige Bibliotheken haben mittlerweile Zeitschriften im Angebot, deren Gesamtzahl sich in der Größenordnung 10.000 bewegt. Eine genaue inhaltliche Orientierung in diesem unübersehbaren Angebot ist weder WissenschaftlerInnen noch BibliothekarInnen möglich. Einer der Wege, dieses Informationsangebot zu bewältigen, ist der Gebrauch von bibliographischen Datenbanken. Diese Datenbanken, teils verlagseigene Angebote, teils fachlich orientiert oder auch interdisziplinär, ermöglichen eine relativ einfache und schnelle Suche über den Inhalt einer Fülle von Zeitschriften. Als derzeit favorisierte Strategie folgt die Zusammenfassung solcher Angebote unter Portalen mit einheitlicher Suchmöglichkeit. Eine Ergänzung (kein Ersatz) solcher Entwicklungen kann von anderer Seite kommen. Eine mathematische Beschreibung des Publikationsprozesses selbst, wie er etwa im Rahmen des Yule-Simon-Prozesses erfolgt und in dieser Arbeit weiterentwickelt wurde, verspricht, wertvolle Erkenntnisse über die Strukturen des Angebots wissenschaftlicher Kommunikation und Publikation zu liefern. Das Bradford Law of Scattering – oder mathematischer: der Exponent der zu einer Stichprobe angepassten Potenzfunktion (oder ein anderes Konzentrations- bzw. Diversitätsmaß) – beschreibt den Grad der Interdisziplinarität eines Fachgebiets. Man könnte sich hier etwa vorstellen, dass, vorausgesetzt es zeigen sich systematische Unter-

schiede in den Exponenten (oder der anderen Maße), die die Verteilungen bezüglich der Stichworte aus verschiedenen Fachgebieten beschreiben, diese Exponenten als wertvolle Eingangsparameter für ein Erwerbungs- oder Etatmodell dienen. Ergäben sich z.B. bei Stichproben für Suchanfragen aus den geisteswissenschaftlichen Bereichen systematisch kleinere Exponenten als für den Bereich Ingenieurwissenschaften, so ließe dies auf eine größere Interdisziplinarität der Geisteswissenschaften schließen. In dem Fall müssten seitens der Bibliothek (natürlich unter Berücksichtigung aller anderen relevanten Punkte wie z.B. Zahl der Lehrstühle, Studentenzahlen etc.) eine relativ gesehen größere Zahl an Zeitschriften aus eben diesem Bereich abboniert bzw. lizenziert werden. Hier läge also eine Möglichkeit, die Strukur der Verteilung der Bibliotheksmittel auf einzelne Fachbereiche den tatsächlichen Strukturen der Publikationslandschaft anzupassen. Dass die Verteilung von Bibliotheksetats derzeit in der Regel durch ganz andere Vorgaben gesteuert wird (in der Regel eher durch die Kräfteverhältnisse der einzelnen Fakultäten als durch objektive bibliothekarisch begründbare Überlegungen), sollte dabei nicht entmutigen. Nur wenn die Bibliotheks- und Informationswissenschaft überzeugende Konzepte bereithält, kann sie darauf hoffen, dass diese langfristig zur Anwendung kommen.

Literaturverzeichnis

[Hinweis:] Für einen umfassenden Überblick über moderne Entwicklungen im Bereich Bibliometrie/Informetrie sei auf das hervorragende Werk von [*Egghe*, 2005] und das darin enthaltene Literaturverzeichnis verwiesen. Der ausführliche mathematische Hintergrund zu Potenzgesetzen inklusive umfangreicher Literaturangaben findet sich bei [*Mitzenmacher*, 2003] und [*Newman*, 2005].

[*Adamic*, 2000] Adamic, L., Huberman, B., The nature of markets in the World Wide Web. *Quarterly Journal of Electronic Commerce* **1**, 512 (2000).

[*Auerbach*, 1913] Auerbach, F., Das Gesetz der Bevölkerungskonzentration. *Petermanns Geographische Mitteilungen* **59**, 74–76 (1913).

[*Ball*, 2004] Ball, R., Tunger, D. Bibliometrische Analysen - ein neues Geschäftsfeld für Bibliotheken?, *B.I.T. online*. **4**, (2004)

[*BMBF*, 2002a] Bundesministerium für Bildung und Forschung, Zukunft der wissenschaftlichen und technischen Information in Deutschland, http://www.bmbf.de/pub/zukunft_der_wti_in_deutschland.pdf, (Letzter Zugriff: 23.1.07)

[*BMBF*, 2002b] Bundesministerium für Bildung und Forschung, Information vernetzen - Wissen aktivieren, http://www.bmbf.de/pub/information_vernetzen-wissen_aktivieren.pdf, (Letzter Zugriff: 23.1.07)

[*Bornholdt*, 2001] Bornholdt, S., Ebel, H., World wide web scaling exponent from Simon's 1955 model, *Physical Review E*. **64**, 035104 (2001)

[*Bradford*, 1934] Bradford, S.C. Sources of Information on Specific Subjects, *Engineering: An Illustrated Weekly Journal*. **137**, 85 (1934)

[*Chen*, 1986] Chen, Y., Leimkuhler, F., A Relationship between Lotka's law, Bradford's law, and Zipf's law, *Journal of the American Society for Information Science*. **37**, 307 (1986)

[*Chen*, 1995] Chen, Y., Chong, P., Tong, M., Dynamic behavior of Bradford's law, *Journal of the American Society for Information Science*. **46**, 370 (1995)

[*Crovella*, 1996] Crovella, M., Bestavros A., Self-similarity in World Wide Web traffic: Evidence and possible causes. In B. E. Gaither and D. A. Reed (eds.), *Proceedings of the 1996 ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, pp. 148–159, Association of Computing Machinery, New York (1996).

[*de Solla Price*, 1963] de Solla Price, D.J., Little science, big science. New York: Columbia University Press (1963)

[*de Solla Price*, 1965] de Solla Price, D.J., Networks of scientific papers. *Science* **149**, 510 (1965).

[*Egghe*, 2005] Egghe, L., Power laws in the information production process: Lotkaian Informetrics. Amsterdamm: Elsevier (2005)

[*Eigenbrodt*, 2006] Eigenbrodt, O., Herausforderung Wissenschaftsgesellschaft, Die Digitale Bibliothek zwischen Mensch, Umwelt und Politik, *In: Vom Wandel der Wissensorganisation im Informationszeitalter - Festschrift für Walther Umstätter zum 65. Geburtstag*. http://edoc.hu-berlin.de/docviews/abstract.php?lang=ger&id=27537, (Letzter Zugriff: 23.1.2007)

[*Estoup*, 1916] Estoup, J., *Gammes Stenographiques*. Paris: Institut Stenographique de France (1916).

[*Fröhlich*, 1999] Fröhlich, G. Das Messen des leicht Meßbaren. Outputindikatoren, Impact-Maße: Artefakte der Szientometrie?, *GMD Report 61*. 27 (1999)

[*Gardiner*, 2004] Gardiner, C., Handbook of stochastic methods, Berlin: Springer (2004)

[*Gorraiz*, 2006] Gorraiz, J., Web of Science versus Scopus oder Das aktuelle Dilemma der Bibliotheken. *Online-Mitteilungen*. **85**, 3, (2006)

[*Gutenberg*, 1944] Gutenberg, B., Richter F., Frequency of earthquakes in california. *Bulletin of the Seismological Society of America* **34**, 185–188 (1944).

[*Heinz*, 2006] Heinz, M., Bemerkungen zur Entwicklung der Internationalität der Forschung - Bibliometrische Untersuchungen am SCI, *In: Vom Wandel der Wissensorganisation im Informationszeitalter - Festschrift für Walther Umstätter zum 65. Geburtstag*. (2006) http://edoc.hu-berlin.de/docviews/abstract.php?lang=ger&id=27537, (Letzter Zugriff: 23.1.2007)

[*Helbing*, 1999] Helbing, D., Schreckenberg, M., Cellular automata simulating experimental properties of traffic flows. *Physical Review E* **59**, R2505-R2508. (1999)

[*Hulme*, 1923] Hulme, E.W. Statistical Bibliography in Relation to the Growth of Modern Civilization, London: Grafton (1923)

[*Levene*, 2006] Levene, M., Trevor, F., Loizou, G., A stochastic model for the evolution of the Web allowing link deletion, *ACM Transactions on Internet Technology*. **6**, 117 (2006)

[*Löffler*, 2005] Löffler, K., Umstätter, W., Wagner-Döbler, R., Einführung in die Katalogkunde - vom Zettelkatalotg zur Suchmaschine, Stuttgart: Hiersemann (2005)

[*Lotka*, 1926] Lotka, A.J. The frequency distribution of scientific productivity, *Journal of the Washington Academy of Sciences*. **16**, 317 (1926)

[*Lu*, 1991] Lu, E., Hamilton, R., Avalanches of the distribution of solar flares. *Astrophysical Journal* **380**, 89–92 (1991).

[*Mainzer*, 1999] Mainzer, K. (Hrsg.) Komplexe Systeme und Nichtlineare Dynamik in Natur und Gesellschaft - Komplexitätsforschung in Deutschland auf dem Weg ins nächste Jahrhundert, Berlin: Springer (1999)

[*Mckiernan*, 2005] Mckiernan, G., Bibliometrics, Cybermetrics, Informatrics, and Scientometrics Sites and Sources *Science and technology libraries*. **2**, 107 (2005)

[*Mitzenmacher*, 2003] Mitzenmacher, M. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*. **1**, 226 (2003)

[*Nacke*, 1979] Nacke, O. Informetrie: ein neuer Name für eine neue Disziplin, *NfD*. **6**, 219 (1979)

[*Naumann*, 2006] Naumann, U., Irrläufer einer missverstandenen Szientometrie, *In: Vom Wandel der Wissensorganisation im Informationszeitalter - Festschrift für Walther Umstätter zum 65. Geburtstag*. (2006) http://edoc.huberlin.de/docviews/abstract.php?lang=ger&id=27539, (Letzter Zugriff: 23.1.2007)

[*Neukum*, 1994] Neukum, G., Ivanov, B., Crater size distributions and impact probabilities on Earth from lunar, terrestial-planet, and asteroid cratering data, In: T. Gehrels (ed.), *Hazards Due to Comets and Asteroids*, pp. 359–416, University of Arizona Press, Tucson, AZ (1994).

[*Newman*, 2005] Newman, M.E.J. Power laws, Pareto distributions and Zipf's law *Contemporary Physics*. **46**, 323 (2005)

[*Otlet*, 1934] Otlet, P. Traité de documentation : Le livre sur le livre ; Théorie et pratique, Bruxelles (1934)

[*Pareto*, 1896] Pareto, V., *Cours d'Economie Politique*, Geneva: Droz (1896).

[*Pennock*, 2002] Pennock, D., Flake g.; Lawrence, S., Glover, E., Giles, C., Winners don't take all: Characterizing the competition for links on the web, *Proceedings of the National Academy of Sciences*. **99**, 5207 (2002)

[*Pipp*, 2006] Pipp, E,. Vergleich der von Scopus bzw. Web of Science erfassten Zeitschriften. *Online-Mitteilungen*. **85**, 3, (2006)

[*Pritchard*, 1969] Pritchard, A. Statistical bibliography or bibliometrics?, *J. Doc*. **25**, 4, 348 (1969)

[*Ranganathan*, 1969] Ranganathan, S.R. Librametry and it's scope in subject analysis for document finding systems, *Quantification and librametric studies management of transaltion service, Indian Statistical Institute, Document Research training Center*. (1969)

[*Roberts*, 1998] Roberts, D., Turcotte, D., Fractality and self-organized criticality of wars. *Fractals* **6**, 351–357 (1998).

[*Schneider*, 2006] Schneider, K., Scopus - Web of Science: Ein Versuch einer Bewertung aus pharmakognostischer Sicht. *Online-Mitteilungen*. **85**, 21, (2006)

[*Schweitzer*, 2003] Schweitzer, F., Meinungsbildung, Kommunikation und Kooperation aus physikalischer Perspektive, *Physik Journal*. **2**, 57, (2003)

[*Schweitzer*, 1998] Schweitzer, F., Selbstorganisation in der urbanen Strukturbildung, *In: Nichtlineare Dynamik. Instabilität und Strukturbildung in physikalischen System, Tagungsband des Workshops des SFB 185, Riezlern 1997*, Frankfurt/Marburg (1998)

[*Shannon*, 1976] Shannon, C., E., Weaver, W. Mathematische Grundlagen der Informationstheorie. München, Oldenbourg Verlag, (1976)

[*Simon*, 1955] Simon, H. A. On a Class of Skew Distribution Functions *Biometrika*. **42**, 425 (1955)

[*Sornette*, 2003] Sornette, D., *Critical Phenomena in Natural Sciences*, chapter 14, Heidelberg: Springer, 2nd edition (2003).

[*Stein-Arsic*, 2003] Stein-Arsic, M. et. al. Bibliometrische Analysen als Intrument des Bestandsmanagement in Bibliotheken *B.I.T. online*. **4**, (2003)

[*Theil*, 1967] Theil, H. Economics and Information Theory. Amsterdam, North-Holland, (1967)

[*Thelwall*, 2006] Thelwall, M., Ruschenburg, T. Informationswissenschaft - Grundlagen und Forschungsfelder der Webometrie, *Nachrichten für Dokumentation*. **8**, 401 (2006)

[*Umstätter*, 1997] Umstätter, W. Wissenschaft in der Gesellschaft, *Lecture held at the Humboldt University in Berlin 1997*.
http://www.ib.hu-berlin.de/ wumsta/lectn.html (1997) (Letzter Zugriff: 23.1.2007)

[*Umstätter*, 2005] Umstätter, W. Anmerkungen zu Birger Hjorland und Jeppe Nicolaisen: Bradford's Law of Scattering: Ambiguities in the Concept of Subject. *Libreas*, **3**, (2005)

[*Wildner*, 2006] Wildner, P., Web of Science - Scopus: Auf der Suche nach Zitierungen. *Online-Mitteilungen*. **85**, 18, (2006)

[*Wolf*, 1997] Wolf, F. Lorenzkurvendisparität - Neuere Entwicklungen, Erweiterungen und Anwendungen. Frankfurt am Main, Peter Lang, (1997)

[*Zanette*, 2001] Zanette, D., Manrubia, S., Vertical transmission of culture and the distribution of family names. *Physica A* **295**, 1–8 (2001).

[*Zipf*, 1949] Zipf, G., *Human Behaviour and the Principle of Least Effort*. Reading, MA: Addison-Wesley (1949).